Journal of Political Economy

JOURNAL OF
POLITICAL ECONOMY

Edited by

ROBERT J. BARRO HARRY G. JOHNSON
SAM PELTZMAN GEORGE J. STIGLER

In cooperation with OTHER MEMBERS of the DEPARTMENT OF
ECONOMICS and the GRADUATE SCHOOL OF BUSINESS
of the UNIVERSITY OF CHICAGO
AND OUTSIDE REFEREES
(SEE INSIDE BACK COVER)

DIANE WELCH, Editorial Assistant DEBORAH ALIBER, Secretary

The Journal of Political Economy is published bimonthly in February, April, June, August, October, and December by The University of Chicago Press. Subscription rates, U.S.A.: institutions, 1 year $20.00, 2 years $38.00, 3 years $55.00; individuals, 1 year $15.00, 2 years $28.00, 3 years $40.00. Student subscription rate, U.S.A.: 1 year $9.00 (letter from professor must accompany subscription). Other countries add $1.00 for each year's subscription to cover postage. Single copy rates: institutions $4.00, individuals $3.00. Supplements: institutions $4.50, individuals $3.50. Back issues are available from 1962 (vol. 70). Make all remittances payable to *Journal of Political Economy*, The University of Chicago Press, in United States currency or its equivalent.

Claims for missing numbers should be made within the month following the regular month of publication. The publishers expect to supply missing numbers free only when losses have been sustained in transit and when the reserve stock will permit.

Business correspondence should be addressed to The University of Chicago Press, 5801 Ellis Avenue, Chicago, Illinois 60637.

Letters to the editors and manuscripts should be addressed to the Editor of the *Journal of Political Economy*, 1126 East 59th Street, Chicago, Illinois 60637. **Manuscripts should be submitted in duplicate (3 copies if possible), accompanied by a $25.00 submission fee made payable to the Journal.** The fee will be refunded to authors whose papers are accepted for publication in the *Journal*. The proceeds from the submission fees are used to pay for refereeing services. Accepted manuscripts must be typed according to the University of Chicago *Manual of Style*. References should be typed double-spaced at the end of the article. Footnotes should be numbered in sequence and double-spaced following the references. Tables should follow the footnotes. Originals of the figures, drawn in india ink, should be submitted if the manuscript is accepted. Abstracts not exceeding 100 words should be submitted in duplicate along with the manuscript.

Applications for permission to quote from this *Journal* should be addressed to The University of Chicago Press and will be freely granted. Reproduction of articles requires written permission from the publisher and the author.

Reprinted volumes 1–72 available from Walter J. Johnson, Inc., 355 Chestnut Street, Norwood, New Jersey 07648. Volumes available **in microfilm** from University Microfilms, 300 North Zeeb Road, Ann Arbor, Michigan 48106; **in microfiche** from J. S. Canner & Co., 49–65 Lansdowne Street, Boston, Massachusetts 02215.

Notice to subscribers: If you change your address, please notify us and your local postmaster immediately, giving *both* your old and your new address. *Allow four weeks for the change.*

Second-class postage paid at Chicago, Illinois, and at additional mailing office.
© 1975 by The University of Chicago. All rights reserved.
ISSN 0022-3808
PRINTED IN U.S.A.

Journal of Political Economy

Volume 83 Number 1 February 1975

Articles

1 Production Technologies of the U.S. Boiler and Turbogenerator Industries and Hedonic Price Indexes for Their Products: A Cost-Function Approach
Makoto Ohta

27 Household Allocation of Time and Church Attendance
Corry Azzi and Ronald Ehrenberg

57 Substitution among Capital, Labor, and Natural Resource Products in American Manufacturing
David Burras Humphrey and J. R. Moroney

83 Capacity Utilization under Alternative Regulatory Restraints: An Analysis of Taxi Markets
Arthur S. De Vany

95 An Estimate of the Liquidity Problem
J. Huston McCulloch

121 Trade in Technological Knowledge and the National Advantage
Carlos Alfredo Rodriguez

137 Full-Time Schooling in Life-Cycle Models of Human Capital Accumulation
T. D. Wallace and L. A. Ihnen

157 Some Problems in the Use of Income-contingent Loans for the Finance of Higher Education
Marc Nerlove

Communications

185 The Quantity Theory in the Early Monetary System of West Africa with Particular Emphasis on Nigeria, 1850–1895
Okonkwo A. Nwani

195 The Equilibrium Size of a Budget-maximizing Bureau: A Note on Niskanen's Theory of Bureaucracy
Albert Breton and Ronald Wintrobe

209 Property Rights and the Dynamic Inefficiency of Capitalism: Comment
H. E. Frech III

215 The Interest Elasticity of Savings, Income Taxes, and the Permanent-Income Hypothesis
Paul J. Taubman

Miscellany

219 The Economics of Reducing Faculty Teaching Loads: Comment
Douglas Needham

225 Assistant Professors Should Be Discriminated Against, or The Less Productive I Am, the More I Should Be Paid
R. P. Parks

Puzzles and Problems

227 *Douglas D. Purvis*

Book Reviews

231 Charles P. Kindleberger, *The World in Depression, 1929–1939*
Anna J. Schwartz

238 John L. Palmer, *Inflation, Unemployment and Poverty*
Anthony M. Santomero

Journal of Political Economy

Volume 83 Number 2 April 1975

Articles

241 "Rational" Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule
Thomas J. Sargent and Neil Wallace

255 The Relevance of the Household Production Function and Its Implications for the Allocation of Time
Robert A. Pollak and Michael L. Wachter

279 Capital-Skill Complementarity, Income Distribution, and Output Accounting
P. R. Fallon and P. R. G. Layard

303 On Two Specifications of Asset Equilibrium in Macroeconomic Models
Duncan K. Foley

325 Covered Interest Arbitrage: Unexploited Profits?
Jacob A. Frenkel and Richard M. Levich

339 Strategic Considerations in the Richardson Model of Arms Races
Michael D. Intriligator

355 Unemployment, Underemployment, and Optimal Job Search
B. Curtis Eaton and Philip A. Neher

377 Collective Choice and General Fund Financing
Edgar K. Browning

Communications

391 Keynes and the Economic Activists of Pre-Hitler Germany
George Garvy

407 More on Windfall Income and Consumption
Tong Hun Lee

419 Benefit-Cost Shadow Pricing in Open Economies: An Alternative Approach
Robin Boadway

431 On the Indirect Productivity of Money
Emil-Maria Claassen

Book Reviews

437 L. P. Carpenter, *G. D. H. Cole: An Intellectual Biography*
T. W. Hutchison

438 J. Philip Cooper, *Development of the Monetary Sector: Prediction and Policy Analysis in the FRB-MIT-Penn Model*
Arthur J. Rolnick

441 Roger L. Miller and Raburn M. Williams, *Unemployment and Inflation: The New Economics of the Wage-Price Spiral*
Jeremy J. Siegel

444 Jacob Mincer, *Schooling, Experience, and Earnings*
Robert E. Hall

Journal of Political Economy

Volume 83 Number 3 June 1975

Articles

447 Distribution Effects and the Aggregate Consumption Function
Alan S. Blinder

477 The Citation Practices of Doctorates in Economics
George J. Stigler and Claire Friedland

509 The Demand for Index Bonds
Stanley Fischer

535 Demography, Technology, and Higher Education: Toward a Formal Model of Educational Adaptation
Stephen P. Dresch

571 The Production and Distribution of Cognitive Skills within Schools
Byron W. Brown and Daniel H. Saks

595 Nonmonetary Factors in the Demand for Medical Services: Some Empirical Evidence
Jan Paul Acton

615 Wealth Effect on the Demand for Money
Vikas Chitre

Communications

633 On the General Equilibrium Analysis of Tax Incidence
J. Gregory Ballentine and Ibrahim Eris

645 Peak-Load Pricing under Regulatory Constraint: A Proof of Inefficiency
Leonard Waverman

655 "An Evaluation of Consumer Protection Legislation: The 1962 Drug Amendments": A Comment
Thomas McGuire, Richard Nelson, and Thomas Spavins

663 "An Evaluation of Consumer Protection Legislation: The 1962 Drug Amendments": A Reply
Sam Peltzman

Book Reviews

669 Albert Breton, *The Economic Theory of Representative Government*
William H. Riker

670 Ajit K. Dasgupta and D. W. Pearce, *Cost-Benefit Analysis: Theory and Practice*
J. N. Wolfe, Ed., *Cost Benefit and Cost Effectiveness: Studies and Analysis*
Daniel Wisecarver

671 John Hicks, *The Crisis in Keynesian Economics*
Harry G. Johnson

673 Edward Pessen, *Riches, Class, and Power before the Civil War*
Alice Hanson Jones

Production Technologies of the U.S. Boiler and Turbogenerator Industries and Hedonic Price Indexes for Their Products: A Cost-Function Approach

---

Makoto Ohta

Tohoku University

The hedonic approach has been proposed and used as a method to obtain quality-adjusted (hedonic) price indexes for the durable goods. The previous hedonic studies did not seem to pay careful attention to the economic factors which determine the functional forms of hedonic regression equations. This paper tries to analyze those factors for the U.S. boilers and the turbogenerators purchased by the U.S. steam electric power industry. This goal amounts to analyzing the production technologies of the industries producing these products under the maintained hypothesis of markup pricing. After analyzing them, we obtain the hedonic price indexes.

I. Introduction

There are three criteria for choosing a specific functional form of a regression equation: (1) choosing the most simple convenient form to obtain the estimates of interesting parameters, (2) choosing the form which yields the best fit, and (3) choosing the form which is based on assumptions that are shown to be plausible in the actual world. Criteria (1) and (2) are used in the previous hedonic studies. This paper examines the functional form of the hedonic regression equation from the viewpoint of (1) and (3) instead of (2). This amounts to testing the assumptions about the production technology of the industry producing the good in question under the maintained hypothesis of markup pricing. We derive its hedonic price index from the cost function for producing it.

This paper is based on my Ph.D. thesis, "Hedonic Price Index for Boiler and Turbogenerator: A Cost Function Approach," submitted to the University of California (Berkeley), June 1971. I am most grateful to D. L. McFadden for introducing me to this area. I owe much to D. W. Gaskins, C. B. McGuire, and, especially, Z. Griliches for their valuable suggestions and encouragement.

[*Journal of Political Economy*, 1975, vol. 83, no. 1]

© 1975 by The University of Chicago. All rights reserved.

Section II will point out some possible specification problems involved in the previous hedonic studies from the viewpoint of (3) to motivate our cost-function approach.[1] Section III will extend the orthodox theory of price index to the situation of quality change to prepare for the discussion in Section IV. Section IV will discuss some theoretical topics on the hedonic price index based on the cost function. Section V will present empirical results for the U.S. boiler and turbogenerator industries and their products.

II. Possible Specification Problems Involved in the Previous Hedonic Studies

The fundamental maintained hypotheses of the previous hedonic studies are: (1) price (i.e., value index) is "separable" so that it is the product of the quality-adjusted price index (hedonic price index) and the quality index. Let $P_t(y)$ be the market price at time t of the good with level of physical characteristics y; $\bar{P}_t$ be the hedonic price index at time t, which does not depend on y; and $q(y)$ be the quality of the good with level of characteristics y. Then, this hypothesis of separability gives us the equation:[2]

$$\underset{\substack{\text{(value}\\\text{index)}}}{\bar{P}_t(y)} = \underset{\substack{\text{(hed-}\\\text{onic}\\\text{price}\\\text{index)}}}{\bar{P}_t}\underset{\substack{\text{(qual-}\\\text{ity}\\\text{index)}}}{q(y)}. \tag{1}$$

(2) The quality of the good $q(y)$ is determined by the consumer's utility function $u(y)$ defined on the physical characteristics of the good y.[3]
(3) Theoretically speaking, we can adopt any "arbitrary" functional form for the quality index q. Griliches (1961) says, "There is no a priori reason to expect price and quality to be related in any particular fixed fashion." He uses a semilogarithmic form for (1), where y_i is the level of the ith characteristic (ith element of y)($i = 1, 2, \ldots, m$):

$$\log P_t(y) = \log \bar{P}_t + a_1 y_1 + \cdots + a_m y_m + \text{disturbance}. \tag{2}$$

He says, "This choice was based on an inspection of the data and the convenience of this particular formulation."

We shall point out below possible specification problems involved in the previous hedonic studies.[4] Price is determined at the intersection of the demand and the supply curves. The functional forms of the demand and the supply curves are not arbitrary. Therefore, the form of the hedonic regression equation is not arbitrary. For example, equation (2) assumes the following hypotheses that may not be satisfied in the actual world. We should test their validity. (*a*) It assumes that the indifference curves of the consumer are parallel. This assumes homotheticity of the utility function and a constant marginal rate of substitution along the indifference curve. (*b*) It assumes that the imputed prices of physical characteristics at a given point in time are constants which do not depend on the level of characteristics. This holds if both indifference curves and expenditure curves are homothetic. (*c*) Equation (2) does not contain cross terms of characteristics like $\sqrt{y_i y_j}$ $(i \neq j)$. This assumes nonjointness of characteristics as outputs in production and utility functions.

---

[1] The previous hedonic studies are reviewed in Griliches (1971). The typical works will be Griliches (1971) and Hall (1969).

[2] $q(y)$ may have a time subscript t. But in order to make a hedonic price index, the functional form of q must be the same at least between the 2 adjacent years. Equation (1) refers to the situation where the time subscript t is, or cannot help being, excluded.

[3] This is often assumed implicitly in the interpretation of the imputed prices of physical characteristics in previous studies. But, to be fair, I should mention that Griliches (1971) attributed the cause of the possible change of the imputed prices over time to the change of supply conditions as well as to taste changes.

The maintained hypothesis (2) of the previous studies considers mainly the consumers. There are only few markets where the price is determined mostly by consumers, such as secondhand markets. Many prices may actually be set by firms according to the markup pricing rather than reflecting consumers' opinion directly. The maintained hypothesis (1)—that is, separability of the hedonic price index from the quality index—is satisfied only if the production technology is input-output separable so that we can write

$$\text{Price} = (1 + \gamma_t)C(y, W_t, t) = \underset{\substack{\text{(quality}\\ \text{index)}}}{C_1(y, t)}\underset{\text{(hedonic price index)}}{C_2(W_t, t)}(1 + \gamma_t)$$

under the hypothesis of markup pricing, where γ_t is the markup ratio at time t; $c = C(y, W_t, t)$ is the unit-cost function of producing a good with characteristics y under the input price vector W_t and the production technology represented by time t.

III. Ideal Price Indexes in the Consumption Theory

We assume no taste difference among consumers of the good and over time. The expenditure function is

$$m(u, P_t) = \underset{y}{\text{Min}}\,[P_t(y) \mid u(y) \geq u].$$

The "ideal quality-adjusted (hedonic) price index" of time t with time t_0 as a base time, $I(t)$, is $m(\bar{u}, P_t)/m(\bar{u}, P_{t_0})$. Suppose that the rational consumer attains the utility level $\bar{u}$ and u_t at the base time t_0 and time t, respectively. Then the "expenditure index" of time t is $m(u_t, P_t)/m(\bar{u}, P_{t_0})$. The "ideal quality index" is $m(u_t, P_t)/m(\bar{u}, P_t)$: the good with a vector of physical characteristics on a higher indifference curve is considered to have a better quality. Under the markup pricing, the ideal hedonic price index is $I'(t)$ below:

---

[4] The specification problems that are discussed here are not "practical": they will not affect the hedonic price index. They are interesting only from a "structural" or "theoretical" point of view. The specification problems such as the "model effects" of Griliches are "practical." They affect the hedonic price index (see Ohta and Griliches 1972 for the model effects).

$$I'(t) = \frac{m(\bar{u}, P_t)}{m(\bar{u}, P_{t_0})} = \frac{(1 + \gamma_t)\ \mathrm{Min}_y[C(y, W_t, t) \mid u(y) \geq \bar{u}]}{(1 + \gamma_{t_0})\ \mathrm{Min}_y[C(y, W_{t_0}, t_0) \mid u(y) \geq \bar{u}]} .$$

IV. Cost-Function Approach

Our cost-function approach to hedonic price indexes is devised to consider the specification problems discussed in Section II and is briefly as follows. We assume markup pricing as a maintained hypothesis so that we have only to consider the supply function, ignoring the demand side.[5] Since the functional form of the supply curve depends on the production technology, our approach tests various hypotheses about the production technology of the industry and then calculates the hedonic price index using the cost function based on these tests.

Definitions of the Indexes

The "total price index" between the good with the vector of physical characteristics $\bar{y}$ produced at time t and the good with y produced at the base time t_0 is the ratio of those prices, $P_t(\bar{y})/P_{t_0}(y)$. Under the hypothesis of markup pricing, this is equal to

$$\frac{(1 + \gamma_t)C(\bar{y}, W_t, t)}{(1 + \gamma_{t_0})C(y, W_{t_0}, t_0)} .$$

The "hedonic price index," $I''(t)$, is the ratio of the price at time t to the price at the base time t_0 which is necessary to obtain the good with the same characteristics $\bar{y}$, $P_t(\bar{y})/P_{t_0}(\bar{y})$. Under the hypothesis of markup pricing, this is equal to

$$\frac{(1 + \gamma_t)C(\bar{y}, W_t, t)}{(1 + \gamma_{t_0})C(\bar{y}, W_{t_0}, t_0)} .$$

---

[5] It is not very unrealistic to assume that markup pricing prevails in the durable goods markets such as automobiles, boilers, and turbogenerators. The coefficients of physical characteristics are interpreted as their imputed costs. The consumers' evaluations of physical characteristics are reflected in the levels of characteristics. If consumers evaluate a characteristic high and its unit production cost is not high, then its level will be large. In other words, the levels of characteristics are determined by the consumers' evaluations and their production costs.

The "quality index" between the good with characteristics $\hat{y}$ and the good with y is the ratio of the costs of obtaining the good with $\hat{y}$ and of obtaining the good with y under the same production technology and input prices, $C(\hat{y}, W_{t_0}, t_0)/C(y, W_{t_0}, t_0)$. Goods with physical characteristics on the higher isocost curves are considered to have higher qualities.

Some Theoretical Topics

First, our indexes satisfy the "factor reversal test": the total price index is equal to the product of the hedonic price index and the quality index.[6] They also satisfy the "time reversal test" if the production technology is input-output separable so that the cost function can be written as

$$c = C_1(y, t)C_2(W_t, t) : I''_y(t_0, t)I''_{\hat{y}}(t, t_0) = 1,$$

where $I''_y(k, l) = P_l(y)/P_k(y)$.

A second topic concerns our quality index. According to the hedonic approach, the quality of a good is measured by the vector of its physical characteristics, y. Ideally, our quality index depends solely on y. But, it depends on input prices and the production technology. If input prices change, the quality index between the good with characteristics $\hat{y}$ and the good with y will change, even though the characteristics y and $\hat{y}$ are held constant. The following extreme case may occur: the good with characteristics $\hat{y}$ has a higher quality than the good with y according to some input prices, but the former has a lower quality according to other input prices.[7] A condition for the quality index not to depend on input prices is the input-output separability of production technology: $c = C_1(y, t)C_2(W_t, t)$. If this is satisfied, the quality index between the good with characteristics $\hat{y}$ and the good with y becomes $C_1(\hat{y}, t_0)/C_1(y, t_0)$ which does not depend on input prices. A similar pathological phenomenon may arise due to technical change. This is avoided if the hypothesis of neutrality of output augmenting technical progress as well as the input-output separability is satisfied so that we can write $C_1(y, t) = C_3(y)C_4(t)$. $C_3(y)$ represents a quality index. Since the form of $C_3(y)$ does not depend on time, we can use a fixed-weight system such as a

---

[6] Our indexes always satisfy this test because the hedonic price index uses "current" y and the quality index uses "base-period" input prices and technology. When the price index uses "base-period" y and the quality index uses "base-period" input prices and technology, the test is satisfied if the production technology is input-output separable.

[7] For example, let the cost function be $c = W_{t1}y_1 + W_{t2}y_2$, where W_{ti} is the ith element of W_t. Let $\hat{y} = (1, 2), y = (2, 1)$, $W_t = (1, 2)$, and $\tilde{W}_t = (2, 1)$. Then the good with characteristics $\hat{y}$ has a higher quality than the good with characteristics y at input price W_t, but it has a lower quality at $\tilde{W}_t$.

Laspeyres or Paasche type.[8] Under these assumptions, we have:

$$P_t(y) = (1 + \gamma_t)\underset{\text{(hedonic price index)}}{C_2(W_t, t)}\underset{\text{(quality index)}}{C_4(t)C_3(y)}.$$

A third topic is that our hedonic price index, $I''(t)$, is not generally equal to the ideal hedonic price index in the orthodox theory, $I'(t)$. $I''(t) \geq I'(t)$, namely, $I''(t)$ has an upward bias as an estimate of $I'(t)$, if physical characteristics $\bar{y}$ solve

$$\underset{y}{\text{Min}}\,[C(y, W_{t_0}, t_0) \mid u(y) \geq \bar{u}].^{9}$$

$I''(t) \leq I'(t)$, if $\bar{y}$ solves

$$\underset{y}{\text{Min}}\,[C(y, W_t, t) \mid u(y) \geq \bar{u}].$$

It is obvious from the definitions of $I'(t)$ and $I''(t)$ that $I'(t) = I''(t)$ when there is no taste change of consumers and no change in the shape of isoexpenditure curves for the good. The shape of the isoexpenditure curve remains unchanged if the following conditions are satisfied under the maintained hypothesis of markup pricing: the input-output separability of the production technology, the neutrality of output-augmenting technical progress, and the output homotheticity concerning physical characteristics so that we can get all the isocost curves for characteristics if we multiply some isocost curve by some appropriate constants.

A final topic is that there are two methods for calculating the hedonic price index in the cost-function approach. One is the method used in the previous studies. We call this the "familiar method." Suppose that the cost function is of Cobb-Douglas type, so that

$$P_t(y) = (1 + \gamma_t)\underset{\text{(hedonic price index)}}{Ae^{-at}\bar{W}_t}\underset{\text{(quality index)}}{y_1^{a_1} \cdot \cdots \cdot y_m^{a_m}},$$

where $\bar{W}_t$ is the total input price index, that is, a weighted average of the individual input prices. Let a caret denote the estimate. Then, the

---

[8] Generally the Laspeyres or Paasche index involves a bias. This condition is important for extending the cross-section regressions of previous hedonic studies to regressions in a time-series framework.

[9] Under the assumption

$$C(\bar{y}, W_{t_0}, t_0)(1 + \gamma_{t_0}) = \underset{y}{\text{Min}}\,[C(y, W_{t_0}, t_0)(1 + \gamma_{t_0}) \mid u(y) \geq \bar{u}],$$

the denominators of $I'(t)$ and $I''(t)$ are the same. Since we do not admit substitution among characteristics y to maintain the utility level $\bar{u}$ but keep the fixed $\bar{y}$ in $I''(t)$,

$$C(\bar{y}, W_t, t)(1 + \gamma_t) \geq \underset{y}{\text{Min}}\,[C(y, W_t, t)(1 + \gamma_t) \mid u(y) \geq \bar{u}].$$

So the numerator of $I''(t)$ is generally larger than that of $I'(t)$. Therefore, $I''(t) \geq I'(t)$. The cause of this upward bias is the same as that of the bias of the Laspeyres price index in the orthodox theory.

hedonic price index is obtained by $P_t(y)/y_1^{\hat{a}_1}\cdot\cdots\cdot y_m^{\hat{a}_m}$. The other is a new method. The movement of the prices over time is decomposed into its components, that is, (*a*) rate of technical change, (*b*) change in the input prices, (*c*) change in the markup ratio. We can compose the hedonic price index from the estimates of these three components:

$$\text{hedonic price index} = \frac{1 + \hat{\vartheta}_t}{1 + \hat{\vartheta}_{t_0}} \cdot \frac{e^{-\hat{a}t}}{e^{-\hat{a}t_0}} \cdot \frac{\hat{W}_t}{\hat{W}_{t_0}} .$$

We call this method the "input cost method," following Gordon (1975).

V. Empirical Results for U.S. Boilers and Turbogenerators: 1947–65

This section will analyze the production technologies and the technical progress of the U.S. boiler and turbogenerator industries and will obtain hedonic price indexes for the industries' products that were purchased by the U.S. steam electric power industry during 1947–65.[10]

The U.S. Boiler and Turbogenerator Industries

According to Edison Electric Institute (1957–59), Combustion Engineering, Inc. (CE) and Babcock & Wilcock Co. (BW) supplied about 80 percent of the total boiler supply to the electric power industry in 1957. Almost all of the turbogenerators were produced by General Electric Company (GE), Westinghouse Electric Corporation (WE), and Allis-Chalmers Manufacturing Company (AC) in 1957–59.

The duopolistic or oligopolistic structure of the industries implies that we may estimate the cost function directly.[11] The reason is as follows. Steam electric power companies are public utilities and must supply the amounts of electricity (Z) that are required by their regions. The variable Z enters the demand function for boilers and turbogenerators as an exogenous variable, but it does not enter the cost function of producing them. The main companies in the industry can be assumed to have very similar (or the same) production technologies and input prices. So the supply curve can be assumed to be fixed except for disturbances at a given point in time. Therefore, direct estimation of the cost function for producing their products will not involve an identification problem. The simultaneous-equation bias will depend on the variance of the disturbance of the supply function. This will be small because boilers and turbogenerators

---

[10] We will mainly discuss the boiler industry; the results of the turbogenerator industry will be presented very briefly.

[11] Since the cost function is a structural equation and the level of the physical characteristic is a decision variable (i.e., endogenous variable) of the buyer of the good, there is generally a simultaneous-equation problem in estimating the cost function directly.

TABLE 1

SAMPLE DISTRIBUTION OF THE BOILER

TECHNICAL PERIOD	BOILER RATING					TOTAL
	0–300	301–600	601–1,000	1,001–1,500	1501–	
1947–50	15	11	3	...	...	29
1951–54	2	12	22	3	...	39
1955–59	1	5	10	10	4	30
1960–65	1	2	2	7	9	21
Total	19	30	37	20	13	119

are expected to be produced under very similar production technologies and input prices at a given point in time.

A price agreement was detected in the turbogenerator industry in 1959, and the judicial decision was delivered by the court in 1961, so the hypothesis of constant markup ratio over time is not plausible in this industry. We assume this hypothesis for the boiler industry because there seems to be no strong evidence against it.

Data and Data Manipulation

Data and Information Sources

1. Dhrymes and Kurz (1964) give us the "technical periods" of the U.S. steam electric power industry, defined as the periods during which the production technology of the industry may be considered to be the same. They are 1947–50, 1951–54, and 1955–59. Further, we treat 1960–65 as a technical period, following Cowing (1969*b*). Since the boiler-turbogenerator complex is the main capital good of the electric power industry, the production technologies of boiler and turbogenerator industries may be considered to be similar or the same within a technical period.

2. Cowing (1969*a*) gives us the following information (the symbol in parentheses is used below). Boilers: (*a*) boiler rating (*BR*)—thousand pounds of steam per hour, (*b*) boiler pressure (*BP*)—psia, (*c*) boiler efficiency (*BE*)—percentage, and (*d*) price (*P*)—tens of dollars. Turbogenerators: (*a*) capacity (z)—megawatt, (*b*) efficiency (*E*)—percentage, and (*c*) price (*P*)—tens of dollars. The sample sizes of boilers and turbogenerators are 119 and 144, respectively. The sample distribution of boilers is shown in table 1. The *BR* and z have a tendency to increase over time. Electric power plant which bought boilers and turbogenerators: (*a*) year or years of initial installation (t); (*b*) type of construction (D_1), conventional (both boilers and turbogenerators are indoor), semioutdoor (turbogenerators are indoor; boilers are outdoor), full outdoor (both

boilers and turbogenerators are outdoor); (*c*) primary fuel (D_2), noncoal, coal; (*d*) primary fuel (D_3). single fuel, multiple fuel; (*e*) number of boilers installed (*NB*), number of turbogenerators installed (*N*); (*f*) average number of maintenance labor per year (*ML*).

3. *Census of Manufactures* (1947, 1954, 1958, 1963) and *Annual Survey of Manufactures* (1947–65) give us the following data about the boiler (SIC 3443) and the turbogenerator industry (SIC 3511); (*a*) cost of materials (thousands of dollars), (*b*) kinds of materials used in the industry, (*c*) cost of each material (thousands of dollars), (*d*) average number of all employees for the year and the total salaries and wages (thousands of dollars), (*e*) average number of production workers for the year, the total man-hours and the total wages (thousands of dollars), (*f*) value of shipments (thousands of dollars).

4. *Wholesale Price Index* (*WPI*) (U.S. Bureau of Labor Statistics 1947–65) gives us the following data: (*a*) WPI of the boiler, the turbogenerator, and the electricity, (*b*) the price indexes of the inputs used in the boiler, the turbogenerator, and the electric power industry.

5. *The Handy-Whitman Index of Public Utility Construction Costs* (Whitman 1947–65) gives us the price indexes of the boilers and the turbogenerators purchased by the U.S. electric power industry.

6. *Fortune's 1966 Input/Output Matrix* (Fortune 1966) gives us the data on: (*a*) the kinds of capital goods used in the boiler (classification 56) and the turbogenerator industry (classification 59), (*b*) the purchase of each capital good by the industry (millions of dollars).

7. *Power* magazine gives us the information on: (*a*) the names of boiler and turbogenerator companies which supplied their products to the electric power industry, (*b*) the total boiler rating of boilers and the capacity of turbogenerators supplied by each company to the electric power industry (1,000 lb). Relative market shares of major companies can be calculated: 55.1 percent and 44.9 percent for CE and BW, respectively in 1957; 50.5, 40.5, and 9.0 percent for GE, WE, and AC, respectively, in 1957–59.

8. *Moody's Industrial Manual* (1947–65) gives us the following data for CE, BW, GE, WE, and AC: (*a*) dividends per share (dollars), (*b*) price range of the share (dollars), (*c*) number of shares, (*d*) depreciation and amortization (dollars), (*e*) property, plant, and equipment (gross) (dollars), (*f*) total income (dollars), and (*g*) federal income tax (dollars).

Compilation of Input-Price Indexes

We classify various inputs of the industries into three groups: labor, materials, and capital goods. The wage index (W_1) of year *t*, W_{1t}, is obtained from the *Census of Manufactures* and *Annual Survey of Manufactures* by dividing the total wage by the man-hours of production workers in year t. Most of the materials used in boiler and turbogenerator industries are included in *WPI* classification 1014 (finished steel product) and 1025 (mill shapes) and *WPI* 1014 and 1015 (foundry and forge shop product), respectively. The price index of materials (W_2) in year t, W_{2t}, is a weighted average of *WPIs* of classification 1014 and 1026 or 1014 and 1015 with their cost shares as a fixed-weight system.[12]

Rental price index of capital service (W_3) is calculated as in Jorgenson and Siebert (1968): the rental price of capital service in year t, v_t, is approximately $q_t(r_t + \delta)/(1 - u)$, where q_t is the price of investment goods, r_t is the cost of capital in year t, and u and δ are explained in footnote 12. The variable r_t is equal to the sum of $(R_t - D_t)/V_t$ and $\dot{V}_t/V_t$, where R_t is the net cash flow, D_t is the direct taxes on business income, and V_t is the market value of the firm at time t. We calculated r_t, δ, and u for *CE* and *BW* by using the data in *Moody's Industrial Manual.*[13] We averaged $(r_t + \delta)/(1 - u)$ over the whole period (1947–65) because the estimate of r_t is not very reliable and was found to be very unstable over the period. Since we cannot identify the producer of the boiler in Cowing (1969*a*), we make a weighted average of $(r + \delta)/(1 - u)$ of CE and BW with their relative market shares as weights. This value is 0.2343. Similarly, it is 0.2076 for the turbogenerator industry. The price index of investment goods (i.e., the index of q) is a weighted average of WPIs of the major investment goods used in the industries with their relative shares of the purchases as weights. The shares of the purchases of various investment goods are calculated from *Fortune's 1966 Input/Output Matrix* (1966).

Total price index ($\bar{W}$) is a weighted average of W_1, W_2, and W_3 with the average cost shares of labor, materials, and capital services in 1962, 1963, and 1964 as a weight system.[14]

---

[12] Their cost shares (average of 1958 and 1963 shares) are 94.4 percent and 5.6 percent or 30.6 percent and 69.4 percent, respectively.

[13] The annual increase of property, plant and equipment, and the total income are used for the investment and sales minus the value of current inputs, respectively. R_t is equal to sales minus the value of current inputs minus investments. V_t is obtained as the product of the mean price of its share and the number of shares. Federal income tax is used for D_t. r_t is obtained from R_t, V_t, and D_t. δ is obtained by averaging the annual depreciation rate (the ratio of depreciation and amortization to the value of property, plant and equipment). u is obtained by averaging the ratio of D_t to total income over the observation period.

[14] The *Annual Survey of Manufactures* (U.S. Department of Commerce 1947–65) provides us with the data for the total wage and the cost of materials in 1962, 1963, and 1964. The rental cost of capital service is estimated by $(r + \delta)K/(1 - u)$, where K is the capital stock in dollar value and is given in the survey. Average cost shares of labor, materials, and capital services in those years are 32.2, 58.3, and 9.5 percent, respectively, in the boiler industry, and 41.8, 48.9, and 9.3 percent in the turbo-generator industry. There are two sets of sufficient conditions for using $\bar{W}$ instead of individual input price indexes in estimating the cost function. The first consists of the hypotheses of input-output separability of the production technology and the neutrality of input-saving technical progress. The second is the proportional movement of the individual input price indexes over time.

Specification Analysis[15]

Test 0: test for no "taste" change in the U.S. steam electric power industry—We test the null hypothesis that there is no taste change for the physical characteristics of the boiler-turbogenerator (BTG) complex in the power industry.[16] The main characteristics of the BTG complex are capacity (z) and efficiency (E). Let P_t^e and P_t^f be the prices of electricity and fuel, respectively.[17] Then, $(P_t^f/P_t^e)(z/E)$ fuel cost is required in terms of electricity to obtain z units of electricity. The net output or the profit in terms of electricity is $z[1 - P_t^f/(P_t^e E)]$. If the power industry maximizes profits, then the industry will be indifferent to the BTG complex which gives the same profit. This gives us the indifference curve of the power industry for z and E. If P_t^f/P_t^e is constant over time, then the indifference curve does not change over time. Thus, we test the null hypothesis of the constancy of P_t^f/P_t^e over time.[18] The unconstrained regression equation is:

$$\frac{P_t^f}{P_t^e} = \text{constant} + \alpha t + u_t,$$

where t denotes the years (1947–65) and u_t is a disturbance. Let us write the t-statistic of an estimate in parentheses. Then, $\hat{\alpha} = 0.79(6.95)$. The null hypothesis of $\alpha = 0$ is rejected at the 1 percent level: the power industry's evaluation of efficiency, relative to capacity, rises over time.

Test 1: test of no scale effect of total industry output on the individual cost function of producing a boiler.—Since we are interested in the individual cost function for producing an individual commodity (not total industry output), we have to test if we can separate the individual cost function from the influence of the scale of total industry output. We use RV_t (value of shipments divided by the Handy-Whitman index of boiler at time t) as total industry output. The unconstrained regression is as follows.

$$\log\left[\frac{P_t(BR, BP, BE, D_2, D_3 : D_1, ML, NB)}{\bar{W}_t}\right]$$
$$= \text{constant} - at + \alpha \log BR + \beta \log BP + \gamma \log BE$$
$$+ \sum_{i=1}^{3} \mu_i D_i + \delta \log ML + \varepsilon \log NB + \phi \log RV_t + u.$$

---

[15] Throughout the following tests and estimations, the disturbance of the regression equation is assumed to satisfy the relevant conditions for the tests and the ordinary least-squares method.

[16] This is one of the conditions for $I^e(t) = I'(t)$.

[17] Labor input is ignored because it is determined largely by the control equipment rather than by the BTG.

[18] We use *WPI* of classification 054 (electric power) for P_t^e and a weighted average of *WPI* 051 (coal), *WPI* 053 (gas fuel), and *WPI* 056 (crude petroleum and natural gasoline) for P_t^f with their cost shares (64, 27, and 9 percent, respectively, in 1967) as a fixed-weight system.

TABLE 2

BOILER: REGRESSION RESULTS

Right-Hand Variable	Estimated Coefficient	t-Statistic
Constant	−.750	−0.19
Semiannual rate of technical progress (a)	.009	2.37
log (BR)	.701	11.62
log (BP)	.509	4.37
log (BE)	.202	0.22
log (ML)	.157	3.63
log (NB)	−.165	−2.97
D_1	−.094	−2.54
D_2	.291	5.10
D_3	.123	2.51

NOTE.—$\bar{R}^2$ = .919; d (Durbin-Watson statistic) = 1.585; number of observations = 119; SSR (sum of squared residuals) = 6.235.

The null hypothesis is $H_0; \phi = 0$; $\hat{\phi} = -.123(-1.05)$. It is not rejected at the 5 percent level, since the critical value is $t_{.025}(106\text{-}11) \approx 2$. We may, therefore, ignore the scale effect of total industry output on the individual cost function. It holds also for the turbogenerator industry.[19]

Estimations 1, 2, and 3: estimations of Cobb-Douglas cost functions.—First, we present the estimation of a Cobb-Douglas cost function with a reliability variable (estimation 1). A reliability variable, $\bar{M}$, of the boiler is assumed to be equal to the maximum boiler rating maintained by one maintenance labor force, BR/ML, adjusted by the type of construction of the electric power plant (D_1) and the number of boilers installed at the plant (NB).

$$\bar{M} = \frac{BR}{ML} e^{-\mu_1' D_1} NB^{-\delta_0'}.$$

Our Cobb-Douglas cost function is

$$\frac{P_t(BR, BP, BE, D_2, D_3, \bar{M} : D_1, NB)}{\bar{W}_t}$$

$$= \text{constant } e^{-at} BR^{\alpha} BP^{\beta} BE^{\gamma} e^{\mu_1'' D_1} e^{\mu_2 D_2 + \mu_3 D_3} NB^{\delta_0''} \bar{M}^{\delta}$$

$$(\delta > 0, \delta_0'' < 0).$$

From these two equations, we get the following equation:

$$\frac{P_t(\cdot, \cdot, \cdot, \cdot, \cdot, \bar{M} : \cdot, \cdot)}{\bar{W}_t}$$

$$= \text{constant } e^{-at} BR^{\alpha + \delta} BP^{\beta} BE^{\gamma} e^{\mu_1 D_1 + \mu_2 D_2 + \mu_3 D_3} ML^{-\delta} NB^{\varepsilon}$$

$$(\mu_1 = \mu_1'' - \mu_1', \varepsilon = \delta_0'' - \delta\delta_0').$$

The regression results are shown in table 2. BR and BP have the highest statistical significance among the three characteristics (BR, BP, and BE). D_2 is highly significant: the cost of producing a boiler with coal as a

---

[19] A possible reason for no scale effect is that both sample boilers and turbogenerators are not made in the factory for market sale but installed by order at the power plant.

TABLE 3

TURBOGENERATOR: REGRESSION RESULTS

Right-Hand Variable	Estimated Coefficient	t-Statistic
Constant	10.157	21.05
Semiannual rate of technical progress (a)	.007	2.80
log (z)	.736	15.99
log (E)	.424	1.52
log (ML)	.049	1.56
log (N)	−.037	−0.89
D_1	−.007	−0.31
D_2	.009	0.24
D_3	.063	2.02

NOTE.—R^2 = .940; d = 1.648; number of observations = 138; SSR = 3.680.

primary fuel is larger than the cost of producing a boiler with noncoal as a primary fuel. $\hat{\delta} = -.157$ has the wrong sign and is statistically significant.[20] The result for the turbogenerator is shown in table 3.[21]

Since the maintenance labor force (ML) has a coefficient with the wrong sign, we did a regression without it (estimation 2).[22] We also estimated the Cobb-Douglas cost function with only the fundamental physical characteristics (BR, BP, and BE for the boiler and z and E for the turbogenerator: estimation 3).[23]

Test 2: test of the cost-minimization hypothesis.—Since this is the basic assumption in using the cost function, we do two tests developed by Hanoch and Rothschild (1970).[24] One uses the yearly quantity data of the industry. The other uses cost and price data.

---

[20] There are four possible explanations for the wrong sign of the coefficient of ML. First, ML is a proxy for something of intrinsic value. This follows Triplett's (1966) explanation of positive and significant coefficient of the weight of a car. Second, we assumed linearity among the logarithms of the variables; but there may be some nonlinear effect. Third, $\hat{M}$ may not be a satisfactory proxy for reliability. Fourth, the wrong sign may reflect a simultaneous-equation bias: what we measure is not a supply curve but a demand curve for reliability.

[21] Unless it is stated otherwise, we assume the markup ratio to be constant over time.

[22] The left-hand variable is log (Price/$\bar{W}_t$) in the estimations of Cobb-Douglas cost functions.

[23] The estimated coefficients of the logarithms of BR, BP, and BE are .755 (10.90), .655(4.45), and 2.520(2.37)(R^2 = .86). Those of z and E are .761(20.43) and .480(1.87) (R^2 = .94).

[24] Suppose that the data set consists of r observations $\{y^i, x^i\}$ on output $y^i > 0$ and n inputs $x^i > 0$. We call the production function $y = F(x)$ "valid" if it is consistent with the observation [i.e., $y^i = F(x^i)$], monotone, quasiconcave, and continuous from above. Let

$$R_i = \{x \mid x \geq x^i\},\ S_i = \{k \mid 1 \leq k \leq r; y^k \geq y^i\},$$

and $T_i = \langle U_{k \in S_i} R_k \rangle$ where $\langle A \rangle$ denotes the convex hull of set A. There exists a valid production function if and only if:

$$x^j \in T_i \Rightarrow y^j \geq y^i \qquad \text{for } 1 \leq i, j \leq r.$$

The data are consistent with a valid production function and the cost-minimization hypothesis if and only if: for $j \ni y^j > y^i$, $W^{i\prime}x^j > W^{i\prime}x^i$ for $1 \leq i \leq r$, where W^i is the ith observation of n input prices. The validity of the production function and the cost-minimization hypothesis are tested by substituting $1/y^i$ for y^i and W^i/c^i for x^i in the above theorem, where c^i is the production cost of the ith observation.

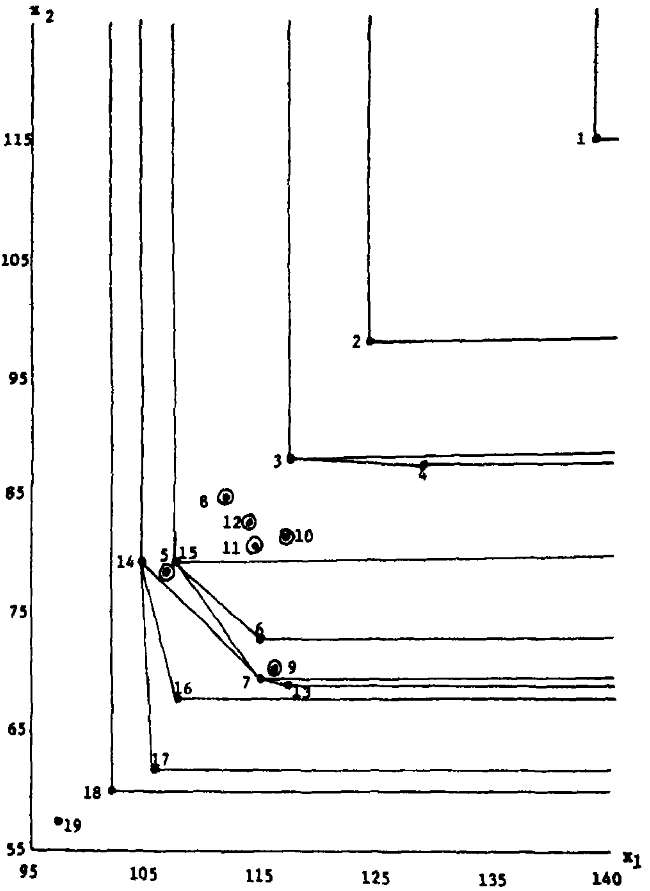

Fig. 1.—Boiler

Test 2-1-1: Hanoch-Rothschild test of the validity of the production function—yearly quantity data.—We use *RV* as output and man-hours of production workers x_1 and real cost of materials (cost of materials divided by W_2) x_2 as inputs.[25] We assume that the production technology is the same over the observation period (1947, 1949–60).[26] Figure 1 plots the observation with x_1 and x_2 as axes, where numbers show the order of magnitudes of *RV*, and the circle (0) around the point (·) shows that the year at the point violates the Hanoch-Rothschild criterion. These violators are 1954, 1957, 1959, 1960, 1961, and 1963. But *RV* of 1954 is almost

---

[25] We do not consider capital services because its cost share is very small in our estimates.

[26] *Annual Survey of Manufactures* (U.S. Department of Commerce 1947–65) lacks the data for 1948.

TABLE 4

BOILER: HANOCH-ROTHSCHILD TEST

Year t	No. Years Whose RV is Larger than RV_t	No. Violations Hanoch-Rothschild Criterion	Smallest Value of Test Statistic (Year)
1947	6	0	...
1949	18	0	...
1950	17	0	...
1951	15	0	...
1952	5	0	...
1953	4	0	...
1954	8	1	.990 (1947)
1955	16	0	...
1956	12	0	...
1957	10	6	.864 (1947)
1958	3	1	.980 (1964)
1959	9	4	.888 (1947)
1960	11	4	.890 (1947)
1961	14	5	.928 (1956)
1962	13	4	.932 (1956)
1963	7	3	.879 (1947)
1964	2	0	...
1965	1	0	...
Total	171	28	...

NOTE.—Since *RV* of 1966 is the largest, 1966 does not appear on the leftmost column.

equal to the *RV*s of 1947 and 1952, and the degree of violation of 1954 is very small. The degree of violation of 1961 is also small. So the serious violators are four out of 19 observations.[27] For the turbogenerator industry, the significant violators are only 2 out of 15 years. Therefore, the hypothesis can be considered to hold in both industries.

Test 2-1-2: Hanoch-Rothschild test of the validity of the production function and the cost-minimization hypothesis—yearly data.—We use $\tilde{W}_1$ (total wage divided by man-hours of production workers) and W_2 as input prices. We form the set of years $\tilde{D}_t$ such that

$$\tilde{D}_t = \{j \mid j \in [1947, 1949\text{–}66];\ RV_j > RV_t\}.$$

Then we calculate the following test statistic for each t:

$$\frac{(\tilde{W}_{1t}x_{1j} + W_{2t}x_{2j})}{(\tilde{W}_{1t}x_{1t} + W_{2t}x_{2t})} \quad \text{for } j \in \tilde{D}_t.$$

If this is smaller than one, the hypothesis is violated. For each year t, we list the number of the violations and the smallest value of the test statistic in table 4. The number of violations (28) is small compared with the number of the test statistics examined (171). Moreover, the degree of violation is small. So, we can assume the cost-minimization hypothesis.

---

[27] The Hanoch-Rothschild test is a strict test which does not allow any stochastic disturbance. We should consider some level of violation to be allowable.

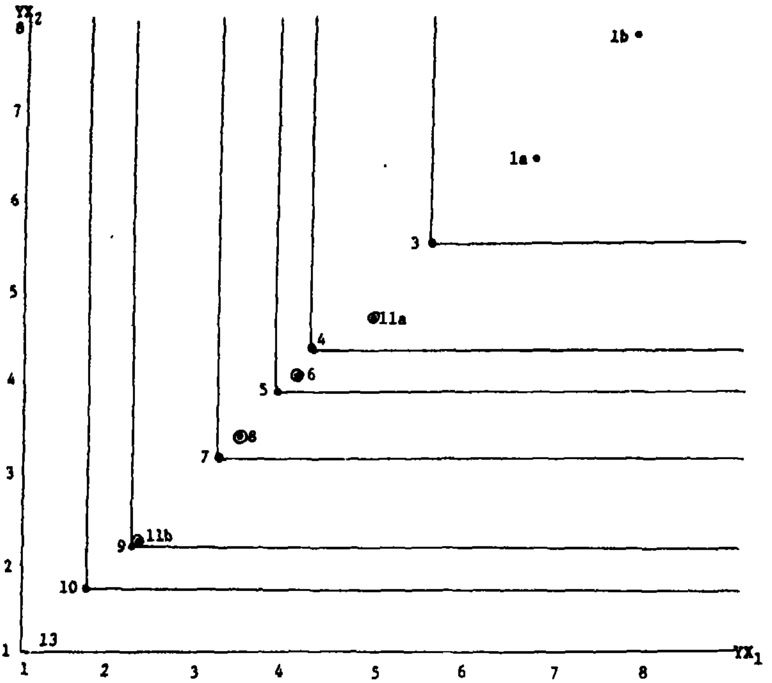

FIG. 2.—Boiler, 1947-50

Test 2-2: Hanoch-Rothschild test of the validity of the production function and the cost-minimization hypothesis: cost and price data.—We transform the data as follows: $YBR = 1{,}000/BR$, $YBP = 1{,}000/BP$, $YX_1 = 1{,}000W_1/P$, and $YX_2 = 1{,}000W_2/P$. In the Dhrymes-Kurz technical period (1947–60), we choose the boiler whose YBP is between 1.0 and 1.2 among the boilers with $D_2 = 1$.[28] There are 13 samples in this technical period. We number them in the order of YBR. The plots of the samples with YX_1 and YX_2 as axes are shown in figure 2. Numbers 6, 8, 11a, and 11b violate the Hanoch-Rothschild criterion. But YX_1's and YX_2's of number 11b and number 9 are almost equal. Thus, the number of serious violations is three out of 13 samples.[29] Therefore, we may assume the cost-minimization hypothesis is valid in 1947–50. This result holds also for other technical periods and for the turbogenerator industry.

Test 3: test of the input-output separability of the production technology.—This is most important for the neutrality of output-augmenting and input-saving technical progresses for the quality index not to depend on input prices. This is also one of the conditions for the price index in the cost-function approach to coincide with that in the orthodox consumption theory and also for using a total input price index.

---

[28] We take BR, BP, and D_2 into consideration because these are the most significant variables in the previous estimations.

[29] The points which violate the criterion buy the boiler with a specified quality at a cheap price. This suggests the possibility that the violating firms are large enough to have large bargaining power. But actually they are not large. This is also the case for other technical periods.

Test 3-1: test of the input-output separability of the production technology—using Diewert (1967) cost function.—We use BR and BP as output measures and W_1 and W_2 as input price indexes.[30] If we let $y_1 = BR$, $y_2 = BP$, and c be the cost, the unconstrained regression is:

$$c = \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{2} \sum_{l=1}^{2} a_{ijkl} \sqrt{y_k y_l} \sqrt{W_i W_j}.$$

The constrained regression is:

$$c = \left(\sum_{k=1}^{2} \sum_{l=1}^{2} \alpha_{kl} \sqrt{y_k y_l} \right) \left(\sum_{i=1}^{2} \sum_{j=1}^{2} \beta ij \sqrt{W_i W_j} \right), \quad \text{where } \beta_{11} = 1.$$

We use the price P as a proxy for c.[31] Since the constrained regression is a linear model with analytic constraints, we use the asymptotic χ^2-test. Let UEC_i be the fitted value for the ith observation using the unconstrained regression, EC_i be the one using the constrained regression, and P_i be the actual price of the ith observation.[32] Then the test statistic is

$$\frac{\sum_{i=1}^{119} (UEC_i - EC_i)^2 119}{\sum_{i=1}^{119} (P_i - UEC_i)^2} = 12.47.$$

Since $\chi^2_{.05}(9 - 5) = 9.49$ and $\chi^2_{.01}(9 - 5) = 13.28$, the hypothesis of input-output separability is rejected at the 5 percent level but not at the 1 percent level.[33] It is not rejected at the 5 percent level for the turbogenerator industry.

Test 3-2: test of the input-output separability of the production technology—using Shephard's lemma.—We use annual data (1947, 1949–66) of the boiler industry. We use RV as output and x_1 and x_2 as inputs.[34] If the cost function is input-output separable so that $c = C_1(RV_t, t)C_2(W_{1t}, W_{2t}, t)$, then by Shephard's lemma,[35]

---

[30] BR and BP are the most significant among the three physical characteristics (BR, BP, and BE). The cost share of W_3 is very small.

[31] Or c times some constant.

[32] The number of observations is 119. $R^2 = .898$ and $.887$, respectively, in the unconstrained and constrained regressions.

[33] This test, however, is not very reliable because the t-statistics of estimates are very low in the unconstrained regression. Let the multiple correlation coefficient adjusted for the degree of freedom be $\bar{R}$. Then $\bar{R}^2 = .894$ in the unconstrained regression and $\bar{R}^2 = .883$ in the constrained regression. So, $\bar{R}^2$ is not very different between the two regressions.

[34] See test 2-1-1 for RV, x_1, and x_2.

$$\frac{x_{1t}}{x_{2t}} = \frac{\partial C(RV_t, W_{1t}, W_{2t}, t)/\partial W_{1t}}{\partial C(RV_t, W_{1t}, W_{2t}, t)/\partial W_{2t}} = \frac{\partial C_2(W_{1t}, W_{2t}, t)/\partial W_{1t}}{\partial C_2(W_{1t}, W_{2t}, t)/\partial W_{2t}}.$$

Therefore, if the hypothesis holds, x_1/x_2 depends only on W_1, W_2, and t: it does not depend on the output level RV_t.[36] The estimated coefficient of log (RV) is not statistically significantly different from zero at the 5 percent level in the boiler industry. It is significantly different from zero at the 5 percent level but not at the 2 percent level in the turbogenerator industry.[37]

Test 4: test of nonjointness of outputs (i.e., physical characteristics).—This hypothesis is assumed in the next test. The unconstrained regression is the same as that of test 3-1. The hypothesis of nonjointness is equivalent to the hypotheses that $a_{ijkl} = 0$ if $k \neq l$. The value of the test statistic is 3.83. The hypothesis is rejected at the 5 percent level but not at the 1 percent level.

Test 5: test of the neutrality of output-augmenting technical progress.—This hypothesis, together with the input-output separability of the production technology, assures us that the functional form of the quality index is fixed over time so that we can use a fixed-weight system in making the quality index. This is also one of the conditions when the price index in the cost-function approach coincides with the index in the orthodox consumption theory. We assume the following cost function:

$$c = \frac{P_t(BR, BP)}{\bar{W}_t} = Ae^{-at}[(BR)^{\alpha} + B(BP)^{\beta}e^{\delta t}],$$

where A and B are constants. δ is a measure of the nonneutrality of output-augmenting technical progress. $\delta = 0$ implies the neutrality. The elasticity of real cost with respect to BR is $\alpha/[1 + B(BP^{\beta}/BR^{\alpha})e^{\delta t}]$.

---

[35] Since we ignored W_3 (the rental price index of capital service), we have to show that it does not affect the labor-materials ratio (x_1/x_2) below. We test this by testing the null hypothesis that $\alpha + \beta = 0$ in the following equation, because x_{1t}/x_{2t} depends on W_{2t}/W_{1t}, input-saving technical progress, and W_{3t}/W_{1t}.

$$\frac{x_{1t}}{x_{2t}} = e^{\text{constant}+at}W_{1t}^{\alpha}W_{2t}^{\beta}.$$

If $\alpha + \beta \neq 0$, it depends not only on W_{2t}/W_{1t} and the input-saving technical progress, but also on something else. From the above discussion, that something must be the effect of W_3. The value of the test statistic is .042 where $F_{.05}(1,15) = 4.54$. So, the hypothesis is not rejected at the 5 percent level. It is not rejected in the turbo-generator industry, either.

[36] If the production technology satisfies the input homotheticity as well as the separability, x_1/x_2 depends only on W_1/W_2 and t.

[37] We tried other output measures, i.e., an annual average boiler pressure and an annual average efficiency for the boiler and the turbogenerator industries, respectively. Their coefficients are not significantly different from zero at the 5 percent level.

So, if $\delta = 0$ (neutral), the elasticity does not depend on t. But if $\delta \neq 0$, it depends on t. Therefore, the hypothesis of neutrality is tested by testing the hypothesis that the elasticities of cost with respect to BR and BP are constant over time.

We use the cost function $c = Ae^{-at}BR^{\alpha}BP^{\beta}$. In the unconstrained regression, α and β can change their values in different Dhrymes-Kurz technical periods, although they are not allowed to do so within the same period. In the constrained regression, α and β do not change over the whole observation period. The value of the test statistic is .24. The hypothesis is not rejected at the 5 percent level [$F_{.05}(6,109) \approx 2.18$]. It is not rejected at the same level for the turbogenerator industry, either.

Test 6: test of the neutrality of input-saving technical progress.—This is one of the conditions for using the total input price index. We use x_1 and x_2 as inputs.[38] We consider the cost function of the total industry output Y, that is, $c_t = C(Y_t, W_{1t}, W_{2t}, t)$. We assume the input-output separability as a maintained hypothesis so that

$$c_t = C_1(Y_t, t)C_2(W_{1t}, W_{2t}, t).$$

Then, by Shephard's lemma, x_{1t}/x_{2t} depends on W_{1t}, W_{2t}, and t. Further, assume the neutrality of input-saving technical progress so that $C_2(W_{1t}, W_{2t}, t) = C_3(W_{1t}, W_{2t})C_4(t)$. Then, by Shephard's lemma, x_{1t}/x_{2t} depends only on W_{1t} and W_{2t}. So the hypothesis of the neutrality can be tested by testing the hypothesis that the coefficient of t is zero in the equation

$$\log\left(\frac{x_{1t}}{x_{2t}}\right) = \text{constant} + at + \alpha \log(W_{1t}) + \beta \log(W_{2t}).$$

The t-statistic of $\hat{a}$ is 1.32. Since $t_{.025}(19 - 4) = 2.131$, the hypothesis is not rejected at the 5 percent level. It is not rejected at the same level in the turbogenerator industry, either.

Test 7: proportionality of the movements of input prices over time.—This is a condition for using the total input price index. We consider only W_1 and W_2.[39] The null hypothesis is that the time coefficient is zero in the regression of input price ratio on constant and time. The t-statistic of the time coefficient is 2.752. The hypothesis is rejected at the 5 percent level but not at the 1 percent level, where critical values are $t_{.025}(19 - 2) = 2.110$ and $t_{.05}(19 - 2) = 2.898$. It is rejected at the 1 percent level for the turbogenerator industry.

---

[38] See text 2-1-1 for x_1 and x_2.

[39] We do not consider W_3 because the cost share of capital services is very small and the estimate of its rental price index W_3 is not very reliable.

Now, the relative input price change is not very large ($\hat{a} = .004$). Further, the elasticity of substitution between inputs is less than one.[40] So, we can guess that the change in the input structure due to the relative input price change has not been large over time. If this guess is correct, a small change in the input price ratio will not change the production cost and the functional form of the cost function. Since this is true for both industries,[41] we can use the total input price index.

Test 8: test of output homotheticity of the production technology.—This is one of the conditions for the price index in the cost-function approach to coincide with the one in the consumption theory. This is also the null hypothesis that the elasticity of cost with respect to the level of the physical characteristic is a constant which does not depend on the level of the characteristic. The unconstrained regression is obtained by taking the logarithm of the following equation:

$$\frac{P_t(BR, BP, BE)}{\bar{W}_t} = Ae^{-at}BR^{\alpha_1+\alpha_2 BR}BP^{\beta_1+\beta_2 BR+\beta_3 BP}BE^{\gamma_1+\gamma_2 BR+\gamma_3 BP+\gamma_4 BE}.$$

The null hypothesis is $H_0:\alpha_2 = \beta_2 = \beta_3 = \gamma_2 = \gamma_3 = \gamma_4 = 0$. The value of the test statistic is 2.68. Since the critical F's are $F_{.05}(6,108) \approx 2.19$ and $F_{.01}(6,108) \approx 2.99$, respectively, the hypothesis is rejected at the 5 percent but not at the 1 percent level. It is not rejected at the 5 percent level for the turbogenerator industry.

Test 9: test of constancy of the rate of technical progress with respect to time.—The unconstrained regression equation is obtained by taking the logarithm of the following equation, where t belongs to the ith Dhrymes-Kurz technical period:

$$\frac{P_t}{\bar{W}_t} = Ae^{-a_i t}BR^{\alpha}BP^{\beta}BE^{\gamma}e^{\Sigma\mu_i D_i}NB^{\delta_0}.$$

Here, a_i is the semiannual rate of technical progress in the ith period ($i = 1, 2, 3, 4$). The a_i can take different values for different periods. In the constrained regression, the rate of technical progress is assumed to be constant over the whole observation period. The value of the test statistic is 1.34, where the critical 5 percent level is $F_{.05}(3,107) \approx 2.70$.

---

[40] $\hat{a} = .941(2.80)$ in the estimation of $\log(x_1/x_2) = \text{constant} + a\log(W_2/W_1) + b\log(RV) + ct$.

[41] We test this in the following way. We use estimation 1 as a constrained regression. The unconstrained regression is obtained by adding the term $(W_1/\bar{W})^{\theta_1}(W_2/\bar{W})^{\theta_2}$ to the right-hand side of this estimation. The supposition that the relative input price change did not affect the production cost is tested by the null hypothesis, $\theta_1 = \theta_2 = 0$. The value of the test statistic is 0.457 and the critical F is $F_{.05}(2,107) \approx 3.09$. The hypothesis is not rejected at the 5 percent level. It is not rejected at the same level for the turbogenerator, either.

Therefore, the null hypothesis is not rejected at the 5 percent level.[42] It is rejected at the 1 percent level for the turbogenerator industry.[43]

Hedonic Price Index for the Boiler

Using the Familiar Method

We use the estimates of Cobb-Douglas cost functions with constant rate of neutral technical progress and the total input price index (specifically, estimations 1, 2, and 3), whose assumptions were tested in the previous specification analysis. We calculated an average, *HI*, of the three hedonic price indexes that are obtained from these estimations. Figure 3 plots *HI*,[44] Handy-Whitman index of the boiler (*HWI*), and *WPI* of the boiler (*WPI* 1072). We notice the following: (1) *HWI* and *WPI* are always increasing over time except for a negligibly small decline. But *HI* sometimes declines by a large amount. This will be due to the small number of observations in some years. In fact, if we connect the points ⊙, *HI* moves in more stable fashion. (2) Before 1956, $HWI < HI$, and after 1957, except for 1958, $HWI > HI$. This may suggest that quality adjustment is too little in *HWI* so that *HWI* overestimates the true quality-adjusted price index, if *HI* is reliable. (3) In all years (1947–65) except for 1964 and 1965, *HI* is larger than *WPI*. In all years, except for 1948, *HWI* is larger than *WPI*. (4) The movement of *HI* is more consistent with that of *HWI* than that of *WPI*.

Using the Input Cost Method

We calculate the following:

$$RTP_t = \frac{100\bar{W}_t}{\bar{W}_0}, \qquad RTC_t = \frac{100EXP(-\hat{a}t)}{EXP(-\hat{a})},$$

where $\hat{a}$ is the estimated rate of technical progress. We use the estimates $\hat{a}$ in estimations 1 and 2 to get $RTC1_t$ and $RTC2_t$, respectively. Under the assumption of the constancy of the markup over time, the hedonic

---

[42] The estimated rate of technical progress becomes larger, however, in later periods, although this phenomenon is not statistically significant. Also, the average level of the physical characteristic tends to increase over time. So, the rate of technical progress will be larger for the higher level of the characteristic.

[43] The rate of technical progress tends to increase over time. The hedonic price index based on the assumption of constant rate of technical progress over time will overestimate the true rate of price inflation.

[44] The point ⊙ of *HI* indicates that the number of observations is greater than or equal to six.

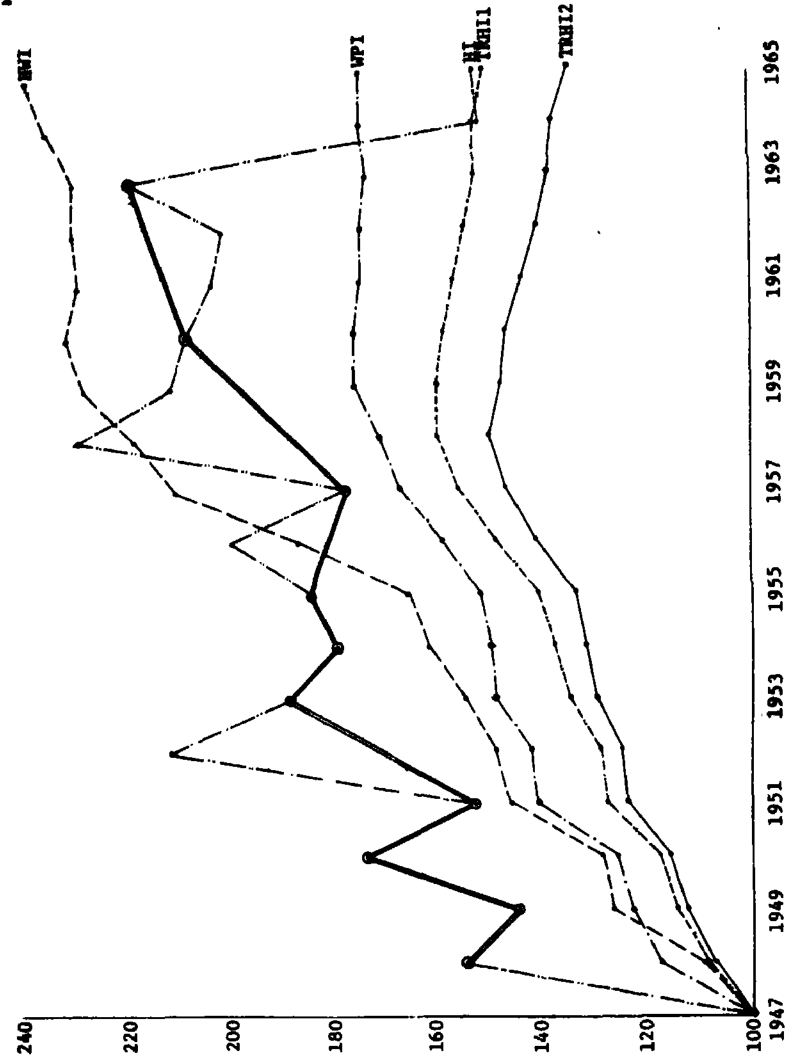

FIG. 3

price index TRHI is:[45] $TRHI_t = RTP_t RTC1_t/100$ or $TRHI2_t = RTC2_t/100$. Figure 3 plots TRHI. We notice the following. (1) In all years, except for 1964, TRHI is smaller than HWI, WPI, and HI. (2) TRHI is most consistent with WPI and is very different from HWI. (3) If everything is measured correctly and the hypothesis of the constancy of the markup ratio over time holds, HI and TRHI should coincide—

---

[45] This will show the trend of the hedonic price index. *RTP*, *RTC*1, *RTC*2, *TRHI*1, and *TRHI*2 are 209, 72, 64, 150, and 134, respectively, in 1965 with 1947 as the base year (i.e., 1947 = 100).

but actually they do not. This may be due to the incorrect measurement of input prices and the invalidity of the above constancy hypothesis.

Hedonic Price Index for the Turbogenerator

Deviation from the Hypothesis of Constancy of the Markup Ratio over Time

We will obtain the estimates of the markup ratio since this industry cannot be assumed to follow the hypothesis. We assume that the Handy-Whitman index at time t, (HWI_t), is proportional to the price of the turbogenerator for fixed z and E. That is,

$$\frac{HWI_t}{HWI_{t_0}} = \frac{P_t(z, E)}{P_{t_0}(z, E)} = \frac{C(z, E, \tilde{W}_t, t)(1 + \gamma_t)}{C(z, E, \tilde{W}_{t_0}, t_0)(1 + \gamma_{t_0})}.$$

Assuming the Cobb-Douglas cost function $c = Ae^{-at}z^{\alpha}E^{\beta}\tilde{W}_t$, we rewrite the above equation as follows:

$$\frac{1 + \gamma_t}{1 + \gamma_{t_0}} = \frac{HWI_t}{HWI_{t_0}} \frac{\tilde{W}_{t_0}}{\tilde{W}_t} e^{a(t-t_0)}.$$

We choose 1950I as the base time, putting $\gamma_{t_0} = 0$ for simplicity. We use the estimated rate of technical progress in estimation 1 for a to calculate the "first estimate of the markup ratio" $\hat{\gamma}_{1t}$.[46] Then, we use $P_t(z, E)/(1 + \hat{\gamma}_{1t})\tilde{W}_t$ as the left-hand variable to obtain the estimated rate of technical progress. Using this estimate, we obtain the "second estimate of the markup ratio" $(\hat{\gamma}_{2t})$. Further, using this estimate, we obtain the "third estimate of the markup ratio" $(\hat{\gamma}_{3t})$.

The above process does not converge to some markup ratio. So it is worthwhile to consider a method for estimating the weight of the physical characteristic in the quality index which does not depend on the estimate of the markup ratio. We assume the following two equations.

$$\frac{HWI_t}{HWI_{t_0}} = \frac{P_t(z, E)}{P_{t_0}(z, E)},$$

$$P_{t_0}(z, E) = A(1 + \gamma_{t_0}) \exp\left(-\sum_{\tau=1}^{t_0} a_\tau\right) z^{\alpha}E^{\beta}\tilde{W}_{t_0},$$

where a_τ is the rate of technical progress at time τ. Then we get (estimation 4):

$$\log P_t(z, E) - \log\left(\tilde{W}_{t_0}\frac{HWI_t}{HWI_{t_0}}\right) = \text{constant} + \alpha \log (z) + \beta \log (E).$$

The estimates are $\hat{\alpha} = .772(21.60)$, $\widehat{\alpha + \beta} = .973(4.44)$, $R^2 = .931$.

---

[46] For example, it is 5.34 percent in 1955I, 31.30 percent in 1959I, and 7.64 percent in 1965I.

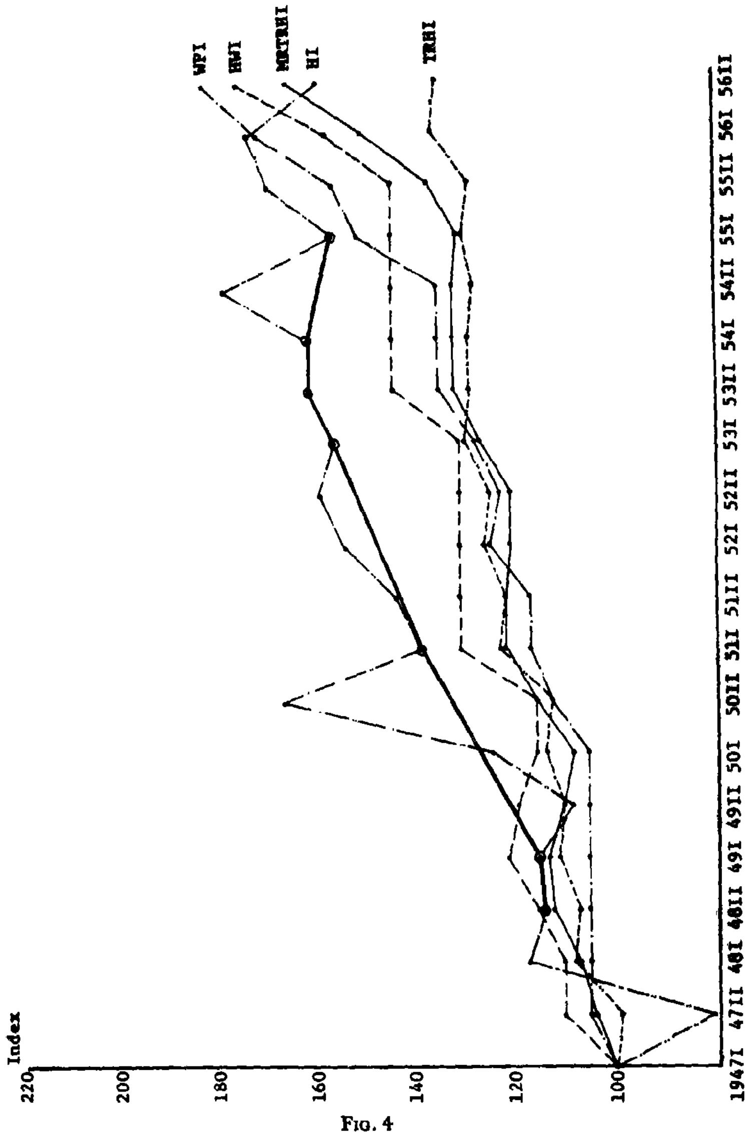

Fig. 4

Using the Familiar Method

Figure 4 plots the average hedonic price index *HI* based on seven regressions. The price agreement was detected in 1959. *WPI*, *HWI*, and *HI* reach their peaks in 1958II, 1959I, and 1960II, respectively. This will be because *HI* is not measured at the order point but at the in-service time point. Since the price agreement seems to have been formed in 1955, *WPI* seems to be more sensitive to the formation of the agreement than *HWI*. The movement of *HI* is more similar to that of *HWI* than to that of *WPI*.

Using the Input Cost Method

We calculate the following: $RTP_t = 100\bar{W}_t/\bar{W}_1$, $RTC_t = 100e^{-\hat{a}t}/e^{-\hat{a}}$, $RMR_t = 100(1 + \hat{\gamma}_t)/(1 + \hat{\gamma}_1)$, $TRHI_t = RTP_t RTC_t/100$, and $MRTRHI_t = TRHI_t RMR_t/100$. We use the first estimate of the markup ratio ($\hat{\gamma}_{1t}$) and $\hat{a}$ in the regression of $\log [P_t(z, E)/(1 + \hat{\gamma}_{1t})\bar{W}_t]$ on constant, time (t), $\log z$ and $\log E$.[47] The input price increased by more than twice during 1947–65. The technical progress reduced the real cost by about 40 percent to produce the same turbogenerator during the period. The net effect is that the trend of the price index (*TRHI*) increased by 40 percent. The markup ratio increased by 5 percent, so the price index increased by about 45 percent over this period. Figure 4 plots *MRTRHI*, *TRHI*, *HI*, *HWI*, and *WPI*.

VI. Summary

First we tested various hypotheses about the production technologies of the U.S. boiler and turbogenerator industries: the physical characteristics of their products were taken into consideration there. We did not find especially strong evidence against the cost-minimization hypothesis and input-output separability of the production technology. This assures us the "existence" of hedonic price indexes and supports our use of simple cost functions. Second, we calculated hedonic price indexes by two methods; the familiar method and the input cost method. The index based on the former increased more over time.

References

Cowing, T. "Firm-Plant Data Sample—U.S. Steam-Electric Generation." Growth Project, Univ. California, Berkeley, 1969. (*a*)

———. "Technical Change in Steam-Electric Generation: An Engineering Approach." Ph.D. dissertation, Univ. California, Berkeley, 1969. (*b*)

Dhrymes, P. J., and Kurz, M. "Technology and Scale in Electricity Generation." *Econometrica*, vol. 35, no. 3 (July 1964).

Diewert, W. E. "A Generalized Leontief Production Function." Working Paper no. 122, Univ. California, Berkeley, C.R.M.S. and I.B.E.R., 1967.

Edison Electric Institute, *Power* (magazine), (1957–59).

Fortune's 1966 Input/Output Matrix.

Gordon, R. J. *The Measurement of Durable Goods Prices.* New York: Nat. Bur. Econ. Res., 1975.

Griliches, Z., ed. *Price Indexes and Quality Change.* Cambridge, Mass.: Harvard Univ. Press, 1971.

---

[47] We tried other estimates for $\hat{\gamma}$ and $\hat{a}$, but *MRTRHI* is not affected much.

Hall, R. E. "The Measurement of Quality Change from Vintage Price Data." Working Paper no. 144, Univ. California, Berkeley, C.R.M.S. and I.B.E.R., 1969.

Hanoch, G., and Rothschild, M. "Testing the Validity of Production and Cost Data." Mimeographed. Harvard Univ., 1970.

Jorgenson, D. W., and Siebert, C. D. "Optimal Capital Accumulation and Corporate Investment Behavior." *J.P.E.* 76, no. 6 (November/December 1968): 1123–51.

Moody's Industrial Manual. New York: Moody's Investor's Service, 1947–65.

Ohta, M., and Griliches, Z. "Makes and Depreciation in the U.S. Passenger Car Market: An Application of the Hedonic Hypothesis to the Construction of Price Indexes and the Study of Market Structure." Mimeographed. Harvard Univ., 1972.

Triplett, J. E. "Measurement of Quality Change." Ph.D. dissertation, Univ. California, Berkeley, 1966.

U.S. Bureau of Census. *Census of Manufactures.* Washington: Government Printing Office, 1947, 1954, 1958, and 1963.

U.S. Bureau of Labor Statistics. *Wholesale Price Index.* Washington: Government Printing Office, 1947–65.

U.S. Department of Commerce. *Annual Survey of Manufactures.* Washington: Government Printing Office, 1947–65.

Whitman, R., and Associates. *The Handy-Whitman Index of Public Utility Construction Costs,* R. Whitman, 1947–65.

Household Allocation of Time and Church Attendance

Corry Azzi

Lawrence University

Ronald Ehrenberg

University of Massachusetts, Amherst

This paper presents the first systematic attempt by economists to analyze the determinants of individuals' participation in religious activities. A multiperiod utility-maximizing model of household behavior is developed which includes among its implications the shape of a household's life-cycle religious-participation profile and the division of religious participation between husband and wife. The theory is empirically tested using statewide church-membership data and survey data on individuals' frequency of church attendance. The paper concludes by discussing several extensions of the model which lead to additional potentially testable hypotheses.

This paper presents the first systematic attempt by economists to analyze the determinants of "religiosity," individuals' religious commitment, and participation in church-related activities. Previous research on the subject conducted by sociologists and psychologists has concentrated on the empirical correlates of religiosity. These studies, using primarily simple correlations or two-way tabular comparisons, have indicated that variables such as race, sex, income, social class, and education tend to be associated

---

Our names are listed in alphabetical order. We are grateful to the numerous individuals who have given us comments on earlier versions of the paper, including Gary Becker, Daniel Hamermesh, Walter Oi, Sherwin Rosen, Theodore W. Schultz, George Stigler, and two referees. None of these individuals is responsible, however, for any errors that remain. An earlier version of the paper was presented to a "Workshop on Applications of Economic Theory" at the University of Chicago, and our concluding section is largely a product of that workshop.
[*Journal of Political Economy*, 1975, vol. 83, no. 1]
© 1975 by The University of Chicago. All rights reserved.

with various dimensions of religiosity.[1] However, a systematic theory of religious participation which simultaneously explains all (or most) of these associations has yet to be developed.[2]

A rapidly growing number of papers by economists has analyzed the allocation of household members' time among various market and nonmarket activities.[3] Recent papers include analyses of the determinants of family size (fertility), household formation (marriage), the value of housewives' time, household nonmarket production, household investments in health, and even suicide rates.[4] Since household members' participation in religious activities requires the use of their time, it seems logical for economists to utilize a similar framework to analyze certain dimensions of religiosity. We utilize such a framework in this paper as the basis for an empirical investigation of variations in church membership and church-attendance rates. Moreover, as we indicate in the concluding section, our framework can be easily extended in several directions to allow one to analyze a wider range of phenomena relating to the interaction of household allocation of time and religiosity.

There is one crucial element, however, that distinguishes an analysis of religious participation from an analysis of participation in other activities. All previous household-allocation-of-time models assume that the expected stream of "benefits" which an individual plans to receive terminates at the time of his death. This assumption is inappropriate for a model of religious participation because most religions promise their members some form of an afterlife. Furthermore, the expected afterlife benefits are often viewed by individuals as being at least partially related to their lifetime allocation of time to religious activities. This suggests that household participation in church-related activities should be analyzed in the context of a multiperiod household-allocation-of-time model which allows for "afterlife consumption," with this variable being at least partially a function of the household's investment of members' time in religious activities during their lifetimes.

In this paper we first provide some evidence on variations in church-attendance rates over time and across individuals and in church-membership rates over areas as well as indicate briefly the empirical correlates of religiosity uncovered by sociologists and psychologists. Next, we outline a multiperiod utility-maximizing model of household behavior which suggests the determinants of household religious participation, including the division of participation between husband and wife and a household's

[1] Examples of this literature are Bultena (1948), Lenski (1953, 1962), Argyle (1959), Lazarwitz (1961), Orbach (1961), Demerath (1965), Glock and Stark (1965), Roof (1972), Sales (1972), and Sklare (1972).

[2] As Roof (1973) indicates, sociologists are well aware of this point.

[3] These papers are based upon the pioneering work of Becker (1965).

[4] See, respectively, Willis (1973), Becker (1973), Gronau (1973), Michael (1973), Grossman (1972), and Hamermesh and Soss (1974).

life-cycle religious-participation profile. While the model is based to a large extent on the salvation motive for religious participation, other motives are introduced, and we indicate how they alter the implications of the model. In the third section, we empirically test our theory using multivariate methods and two different types of data: statewide data on church-membership rates for 1952, 1936, and 1926; and survey data on the frequency of church attendance for 1,500 U.S. adults in 1973. Finally, we suggest several extensions of the model which lead to numerous other potentially testable hypotheses.

I. Background Data and Prior Research

Sociologists have emphasized the multidimensional nature of religiosity, including church membership, church attendance, religious devotionalism, and the extent of orthodox belief. Moreover, empirical studies have indicated that the correlations of the various dimensions with any given socioeconomic variable are not likely to be identical. For example, church attendance has generally been found to be positively associated with measures of social class, while religious devotionalism has been found to be negatively associated. Our research abstracts from the multidimensional nature of religiosity and focuses on church membership and the frequency of church attendance. More generally we are concerned with the household's allocation of time toward religious activities, and we take its members' religious beliefs as predetermined. That is, we attempt to explain the allocation of time to religious activities given religious belief rather than religious belief itself.

Table 1 presents background information on church attendance in the United States. Section A, based on Gallop Poll results, indicates that over the last 30 years church attendance in a given week has fluctuated between 36 and 49 percent of the population, with the fraction of the population attending church in any given week declining steadily since 1958.[5] Unfortunately, the small number of time-series observations makes it impossible to meaningfully test any model of the determinants of church-attendance rates with this national data. Section B indicates, however, that the frequency of church attendance also varies substantially across individuals. We use the cross-section survey data upon which this table is based to test our model in Section III below. Table 2 presents church-membership rates by state for 1926, 1936, and 1952, the last three years in which church-membership figures by state were collected by either a public or private organization. Church-membership statistics are notoriously poor in quality. The number of denominations reporting

[5] Argyle (1959) has presented similar data for the 1900–1937 period in Great Britain. He cautions that individuals tend to overstate their frequency of church attendance.

TABLE 1

A. PERCENTAGE OF AMERICANS WHO ATTEND CHURCH EACH WEEK, SELECTED YEARS 1939–73

Year	%	Year	%	Year	%
1939	41	1957	47	1965	44
1940	37	1958	49	1966	44
1942	36	1959	47	1967	45
1947	45	1960	47	1968	43
1950	39	1961	47	1969	42
1954	47	1962	46	1970	42
1955	49	1963	46	1971	40
1956	46	1964	45	1972	40
				1973	40

NOTE.—Based upon answers to the question, "Did you happen to attend church in the last seven days?" In 1947, the question was asked only of church members.

B. AVERAGE FREQUENCY OF CHURCH ATTENDANCE AND PERCENTAGE OF BELIEVERS IN AN AFTERLIFE, BY RELIGION, 1973

Religion	Sample Size	Average Frequency Religion	Standard Deviation in Annual Frequency of Church Attendance	Believe in Afterlife (%)
All Protestant	940	27.77	31.84	76.06
Baptist	312	29.96	32.23	78.52
Episcopalian	41	15.09	24.02	78.04
Lutheran	124	22.27	25.12	70.97
Methodist	198	22.53	25.09	75.25
Presbyterian	57	27.14	26.80	68.42
Catholic	388	30.63	29.20	70.10
Jewish	42	6.31	13.76	16.67

SOURCES.—A: Historical Gallop Poll data cited in "Worship Habits Found Constant," *New York Times*, January 13, 1974; G. Gallup and J. Davies, *Religion in America* (Princeton, N.J.: American Inst. Public Opinion, 1971); *Special Report on Religion 1967* (Princeton, N.J.: American Inst. Public Opinion, 1967); B. Landis, *Religion in the United States* (New York: Barnes & Noble, 1965). B: Authors' calculations from the National Opinion Research Center's *Spring 1973 General Social Survey* data.

varies over time, different denominations use different criteria for defining what constitutes a member, and systematic underreporting of church members in certain groups or areas has occurred in each census. Consequently, substantial errors in the data are likely. Nevertheless, the table clearly indicates that the interstate variation in church-membership rates was large in each year, with the membership rate varying between roughly 30 and 70 percent. Any economic model of religious participation should contribute to the explanation of this phenomenon.

As noted in the introduction, sociologists have devoted much effort toward ascertaining the empirical correlates of various dimensions of religiosity, primarily using simple correlations or two-way tabular comparisons. In a properly specified multivariate model, not all of these correlations would necessarily be observed. However, it is important to ascertain whether a model of household allocation of time to market and

TABLE 2

CHURCH MEMBERSHIP AS A PERCENTAGE OF POPULATION, BY STATE

	1926	1936	1952
Maine	37.4	37.9	41.2
New Hampshire	49.0	49.4	53.0
Vermont	45.2	47.2	52.7
Massachusetts	61.1	60.8	67.1
Rhode Island	69.1	67.4	75.7
Connecticut	63.1	63.0	60.5
New York	58.1	54.5	60.1
New Jersey	53.8	57.3	57.6
Pennsylvania	56.3	55.3	58.9
Ohio	45.6	43.1	46.4
Indiana	44.4	40.3	43.6
Illinois	46.9	45.7	53.6
Michigan	40.9	35.1	42.3
Wisconsin	52.3	52.5	63.8
Minnesota	51.4	50.1	61.6
Iowa	44.2	43.3	53.6
Missouri	44.7	37.4	48.9
North Dakota	45.7	48.0	63.0
South Dakota	43.9	42.0	58.4
Nebraska	41.7	42.3	53.4
Kansas	40.7	37.7	46.6
Delaware	47.4	44.2	44.3
Maryland	48.7	43.1	47.7
District of Columbia	51.1	45.8	46.6
Virginia	49.3	39.5	38.5
West Virginia	32.8	26.8	32.8
North Carolina	48.1	37.4	39.9
South Carolina	50.9	38.7	40.0
Georgia	46.5	41.6	38.3
Florida	41.7	32.2	36.4
Kentucky	41.5	33.2	44.8
Tennessee	40.6	32.9	40.2
Alabama	48.2	41.3	34.2
Mississippi	41.6	36.9	33.9
Arkansas	34.2	29.8	31.3
Louisiana	52.4	50.3	53.7
Oklahoma	25.8	24.9	42.4
Texas	45.5	37.2	53.6
Montana	28.1	29.1	46.7
Idaho	37.0	36.2	44.1
Wyoming	29.6	28.1	43.6
Colorado	35.4	32.6	41.6
New Mexico	54.2	49.9	64.9
Arizona	38.8	34.8	44.9
Utah	76.3	69.9	73.3
Nevada	23.1	27.2	37.6
Washington	25.9	22.0	30.5
Oregon	26.3	24.1	27.7
California	31.9	30.1	40.7
U.S. average	44.6	41.4	49.2

SOURCES.—*Churches and Church Membership in the U.S.* (New York: Nat. Council Churches of Christ USA, 1956), table 4; U.S. Bureau of the Census, *Religious Bodies 1936* (Washington: Government Printing Office, 1941); *1940 Census of Population: Characteristics of the Population by State* (Washington: Government Printing Office, 1942).

nonmarket activities is capable of generating as implications all (or most) of these previously observed associations. Consequently, we list below some of the observed "regularities" that the sociological literature on the empirical correlates of religiosity has discovered.[6]

1. Women tend to participate in church-related activities more than men.
2. Church attendance tends to increase with age. However, some investigators have noted an initial decline with age until age 30–35, or a decline with age for the extreme elderly. The latter phenomenon is attributed to an increasing incidence of poor health.
3. Only weak relationships are found between income (or social class) and church attendance. Usually the relationship is positive, but some investigators find it backward bending (participation is highest for the middle-income groups).
4. Blacks attend church more frequently than whites.
5. Church attendance is higher in rural areas than urban areas.

II. A Model of Household Religious Participation

Individuals participate in church-related activities for at least three reasons. First, they may view their expected afterlife consumption as being related to their participation; we have previously referred to this as the "salvation motive." Second, they may derive current satisfaction from church membership and activities because of inherent religious beliefs or purely social reasons. We shall refer to this as the "consumption motive." Third, social pressures in a community may suggest that church membership and participation will increase the probability of an individual's succeeding in business. We shall refer to this as the "social-pressure motive," and it will probably be important in areas where only a few established churches exist, especially for new entrants to these communities. Although all three motives are undoubtedly present, our model will primarily consider the salvation motive, as strong testable implications are derived from the resulting framework. However, we shall indicate how the inclusion of the other motives alters the implications of the model.

We consider a household consisting of two members, a husband and a wife. The household is assumed to have a preference ordering that can be denoted by the quasiconcave utility function

$$U = U(C_1, C_2, \dots C_t, \dots C_n, q), \tag{1}$$

---

[6] The list of regularities is derived from the sources cited in n. 1. In all cases the measure of religiosity used in these studies is closely related to either frequency of church attendance or time devoted to church-related activities.

where C_t represents the household's consumption in period t and q the expected value of the household's afterlife consumption.[7] For simplicity, we assume that both household members know the length of their lives and that both will die at the end of period n. Both household members are also assumed to know their current and future market wages, which are taken as predetermined in the model.[8]

Household consumption in period t is given by a household-production function which transforms the household's purchases of a composite market good (x_t) and the time allocation by the husband (h_{1t}) and the wife (h_{2t}) to household consumption into units of the final consumption commodity (C_t). For simplicity, we assume this function to be the same in each period and to be continuously differentiable and concave:

$$C_t = C(x_t, h_{1t}, h_{2t}) \text{ for all } t. \tag{2}$$

Expected afterlife consumption is assumed to be believed by the household to be a continuous differentiable, concave function of the time spent in church-related (or devotional) activities by the husband (r_{1t}) and the wife (r_{2t}) in all periods during their lifetimes:[9]

$$q = q(r_{11}, r_{12}, \ldots r_{1n}, r_{21}, \ldots r_{2n}). \tag{3}$$

---

[7] Our household's view of the afterlife is not one of an all-or-nothing proposition (heaven or hell), it is rather that there exists a continuum of possible outcomes. The stream of expected afterlife benefits will last indefinitely into the future. However, if the household has any positive subjective rate of time preference, we can convert this infinite durational flow into a finite expected discounted present value of afterlife benefits (as of period $n + 1$) which we denote by q. Thus, we can represent an infinite horizon model by an $n + 1$ period model.

[8] The assumption that the time of death is known with certainty will be relaxed shortly. The assumption of exogenous wage implies that our model takes human-capital-investment decisions as predetermined, and it also abstracts from the time allocated to such investments. We consider the implications that can be derived when human-capital decisions are made endogenous in the concluding section of the paper.

[9] This specification assumes that the expected afterlife benefits are shared by the household members and that they *perceive* the allocation of different members' time to religious activities as substitute inputs. The alternative assumption that each household member "produces" his own expected afterlife benefits can be shown to leave unaltered virtually all of the implications which we derive below. Equation (3) also ignores the possibility that individuals may substitute goods (charity) for time in the production of expected afterlife consumption. In the concluding section, we discuss the implications one obtains from allowing this possibility. We should emphasize that eq. (3) need be only *perceived* by the household to exist and need not in fact *actually* exist. One might reasonably question, however, if erroneous perceptions are likely to persist over time. Indeed, George Stigler has orally suggested to us that "the absence of evidence is not completely devoid of evidence." Nevertheless, for our purposes we need assume only that individuals perceive (3) to exist at a point in time. Finally the assumed concavity of the afterlife "production function" is easily justified. If the function were strictly convex, the result would be corner solutions, possibly with all of the household's time spent in religious activities. Individuals who display such behavior are in the minority, are not represented in our empirical data, and would not alter their behavior in response to marginal changes in the exogenous variables of the model.

The precise specification of the function is of course determined by the beliefs of the household members. "Nonbelievers" are simply those individuals who believe that q will always equal zero, regardless of the level of the household members' time inputs. In the context of this model, they would devote no time to religious activities. However, in more general models that allow for the consumption or social-pressure motives, their participation in religious activities would be rational.

Let p denote the price of the market good in any time period, w_{1t} and w_{2t} the wage rates of the husband and wife, respectively, in period t, i a constant market rate of interest, v the household's exogenous nonlabor income in each period, and l_{1t} and l_{2t} the husband's and wife's hours of work in period t, respectively. Assuming that the household plans to leave no estate, its lifetime discounted income constraint is given by

$$\sum_{t=1}^{n} [px_t/(1+i)^{t-1}] = \sum_{t=1}^{n} [(v + w_{1t}l_{1t} + w_{2t}l_{2t})/(1+i)^{t-1}]. \tag{4}$$

If T is the total time available per period, the time constraints it faces are given by

$$h_{jt} + r_{jt} + l_{jt} = T \text{ for } t = 1, 2, \ldots n \quad \text{and} \quad j = 1, 2, \tag{5}$$

where

$$x_t, h_{jt}, r_{jt}, l_{jt} \geq 0 \qquad \text{for all } t \text{ and } j. \tag{6}$$

Statements (1)–(6) comprise a well-defined maximization problem. The production functions (2) and (3) can be substituted into (1) to yield a composite objective function, and the time constraints can be solved explicitly for l_{1t} and l_{2t} and substituted into (4) to yield a "full-wealth" constraint. Together, they allow one to write the Lagrangian function for the problem:[10]

$$\begin{aligned} L = {} & U[C(x_1, h_{11}, h_{21}), C(x_2, h_{12}, h_{22}), \ldots C(x_n, h_{1n}, h_{2n}), \\ & q(r_{11}, \ldots r_{1n}, r_{21}, \ldots r_{2n})] + \lambda\Bigg(\sum_{t=1}^{n} [px_t/(1+i)^{t-1}] \\ & - \sum_{t=1}^{n} \Bigg\{\Bigg[v + \sum_{t=1}^{2} w_{jt}(T - h_{jt} - r_{jt})\Bigg]/(1+i)^{t-1}\Bigg\}\Bigg). \end{aligned} \tag{7}$$

Assuming that an interior solution occurs, the first-order conditions require that at the optimum

$$(\partial q/\partial r_{1t})/(\partial q/\partial r_{2t}) = (w_{1t}/w_{2t}) \text{ for all } t. \tag{8}$$

Suppose the husband and wife are equally efficient in the production of expected afterlife benefits in the sense that, if they both devote the same fraction of their time to religious activities in a period, their "marginal products" in expected afterlife-benefit production will be equal in that period.[11] Equation (8) then implies that if their market wages are equal in the period they will devote the same number of hours to religious activities in the period; however, if the husband's wage exceeds the wife's wage in any period it implies that the wife will devote more hours to religious activities than will the husband. Thus a plausible explanation for women's more frequent participation in church-related activities is simply that they face lower market wages.

---

[10] A mathematical appendix, available from Professor Ehrenberg, formally proves all of the propositions which follow.

If we assume momentarily that both the husband and wife face constant market wage rates over their lifetimes, the first-order conditions can also be shown to require that at the optimum

$$(\partial q/\partial r_{jt})/(\partial q/\partial r_{jt-1}) = (1 + i)^{-1} \text{ for all } t \text{ and } j = 1 \text{ or } 2. \qquad (9)$$

This condition requires that household members reallocate their time toward church-related activities with advancing age. Indeed, if for an individual the marginal product of an additional hour of religious activity is the same in periods $t - 1$ and t when the individual devotes the identical number of hours to devotional activity during the two periods, then equation (9) implies that his number of hours per period devoted to religious activities should increase with age.[12]

The reason for this result should be obvious. Models of optimal accumulation of human capital show that it is optimal for individuals to concentrate their investment in human capital in the early stages of their life cycles since this will enable them to recoup the returns from the investments for a larger number of years. In contrast, the theory of investment in religious capital developed here implies that it is optimal to concentrate "investment" in the latter stages of the life cycle because the expected returns will not be generated until its end. The forces which generate a monotonically declining human-capital-investment profile with age will also generate a monotonically increasing "religious-capital-investment profile" with age.

This conclusion is based upon the assumption that wage rates of all household members are constant over their lifetimes. If for any two adjacent time periods (t and $t - 1$) an individual's wage rates vary, condition (9) becomes

$$(\partial q/\partial r_{jt})/(\partial q/\partial r_{jt-1}) = (w_{jt}/w_{jt-1})(1 + i)^{-1} \quad \text{for } j = 1 \text{ or } 2. \qquad (9')$$

---

[11] Formally, this requires $(\partial q/\partial r_{1t}) = (\partial q/\partial r_{2t})$ when $r_{1t} = r_{2t}$. An example of one production function consistent with this assumption is

$$q = \prod_{t=1}^{n} (r_{1t})^{1/2n}(r_{2t})^{1/2n}.$$

[12] Formally, this requires $(\partial q/\partial r_{jt}) = (\partial q/\partial r_{jt-1})$ when $r_{jt} = r_{jt-1}$. The production function suggested in the preceding footnote is consistent with this assumption.

Consequently, *ceteris paribus*, the more rapid the rate of wage increase, the smaller will be the rate at which hours devoted to religious activity will increase with age. Since age-earnings profiles tend to be steeper for males than females, this condition suggests that age hours devoted to religious-activity profiles will be steeper for females than males. That is, females' hours devoted to religious activities will rise more rapidly with age than will comparable profiles for males because females' marginal cost of investing in religious capital rises less rapidly with age than males'.

If the rate of wage increase between any two periods is sufficiently large, it may "swamp" the effect of discounting and cause religious participation to decline between the two years. Age-earnings profiles tend to be concave, with the rate of increase in earnings declining with age. Consequently, especially for white males who show sharp earnings increase during their twenties, the age–religious-participation profile *may* be U shaped, first declining with age and then increasing, as some sociologists have observed.

The optimality conditions also enable us to calculate the impact on religious participation of a change in the household's exogenous nonlabor income or a proportionate change in an individual's wage rate each period. As long as expected afterlife consumption is a normal good, an increase in the level of exogenous nonlabor income will lead to an increase in religious participation, that is,

$$\partial r_j/\partial v > 0 \qquad \text{for } j = 1 \text{ or } 2. \tag{10}$$

The response of an individual's time devoted to religious activities to a proportionate shift in his wage profile can be decomposed into the sum of an income-compensated own substitution effect $(\partial r_j/\partial w_j)^*$ and an income effect:

$$\partial r_j/\partial w_j = (\partial r_j/\partial w_j)^* + l_j\, \partial r_j/\partial v \qquad \text{for } j = 1 \text{ or } 2. \tag{11}$$

Under weak assumptions, the compensated substitution can be shown to be negative, while the income effect presumably is positive.[13] Thus the uncompensated own substitution effect is unrestricted in sign.

---

[13] The negative sign of the compensated substitution effect follows directly from two facts. First, the opportunity costs of an hour in consumption activity or an hour in religious activity are identical for an individual and equal to his wage rate. A compensated increase in the wage rate must therefore cause a decrease in the sum of nonmarket work hours. Second, religious activity is more time intensive than the consumption commodity, since market goods can be used to produce the latter. Consequently, an increase in the wage rate increases the cost of religious activity relative to the cost of the consumption commodity and induces the individual to substitute consumption for religious activity. Provided that the time utilized in household production increases as current consumption increases, the postulated result follows. Note that if we allowed goods (contributions or the explicit purchases of indulgences) to enter the afterlife-benefit-production function, the sign of the compensated substitution effect would depend upon the relative time intensities of C and q.

Our analysis so far has been restricted to a consideration of only the salvation motive for religious participation.[14] It can be easily generalized, however, to include the consumption motive. That is, if because of either inherent religious beliefs or purely social reasons the household obtains current satisfaction from participating in religious activities, then the household utility function in (1) is replaced by

$$U = U(C_1, s_1, C_2, s_2, \ldots C_n, s_n, q), \tag{1'}$$

where s_t is the consumption value of religious participation in period t. It is assumed to be a continuous concave function of the husband's and wife's times devoted to religious activities during the period:

$$s_t = s_t(r_{1t}, r_{2t}) \quad \text{for all } t. \tag{12}$$

If the husband and wife are equally efficient in the production of both current religious-consumption benefits and expected afterlife benefits, as long as her market wage is lower the wife will continued to spend a greater number of hours in religious activities.[15] Our implications concerning the effects of exogenous nonlabor income and an individual's wage rate on religious participation can also be shown to remain unchanged. In contrast, the generalization of equation (9′) with regard to the intertemporal allocation of time to religious activities is given by

$$\frac{(\partial U/\partial s_t)(\partial s_t/\partial r_{jt}) + (\partial U/\partial q)(\partial q/\partial r_{jt})}{(\partial U/\partial s_{t-1})(\partial s_{t-1}/\partial r_{jt-1}) + (\partial U/\partial q)(\partial q/\partial r_{jt-1})} = \frac{w_{jt}}{w_{jt-1}}(1 + i)^{-1} \quad \text{for all } t \text{ and } j = 1 \text{ or } 2. \tag{9''}$$

Even if an individual faces a constant wage rate during the two periods, there may no longer be any a priori basis for concluding that participation should increase with age because of the consumption motive for religious participation. However, if the salvation motive is substantially more important to individuals than the consumption motive, all of our previous implications concerning age–religious-participation profiles are likely to continue to hold.[16]

---

[14] A referee has pointed out to us that the salvation motive operates in our model in an analogous manner to the bequest motive in models of optimal consumption over time. However, the usual models stress bequests as functions of goods, not of time inputs. The recent work on preschool investment in children, for example Leibowitz (1974), is somewhat similar analytically to our approach in that it involves a bequest of human capital which is created at least partially through time inputs. A second referee has questioned whether the salvation motive is really a major motive for religious participation, citing a study (Croog and Levine 1972) that found that men experiencing a myocardial infarction (heart attack) show virtually no change in their frequency of attendance at religious services over a course of a year following the crisis event. This study was based upon a sample of 324 men and relied primarily on simple tabular comparisons. Rather than argue about the importance of this motive a priori, we prefer to stress the testable implications that it generates and our empirical results which follow.

[15] That is, in addition we now require that $(\partial s_t/\partial r_{1t}) = (\partial s_t/\partial r_{2t})$ when $r_{1t} = r_{2t}$ for all t.

Clearly any factor which increases the current satisfaction which the household receives from participating in religious activities should lead to an increase in the time it allocates to church activities. For example, since market consumption alternatives are relatively limited in rural as opposed to urban areas, the social values of religious activities are probably higher in rural areas, and thus religious participation is higher there. Similarly, racial discrimination may take the form of limiting market consumption alternatives of nonwhites relative to whites and thus lead to higher nonwhite religious participation. In addition, factors such as whether both husband and wife are members of the same religion and the number of school-age children in a family should influence the current consumption value of religious activities and, therefore, the household's religious participation.[17] We should emphasize, though, that observed empirical relationships between the variables discussed in this paragraph and religious participation are not strong tests of the model since they follow solely from the assumption of *systematic variations in preferences*. Rather, they are entered in our empirical analyses primarily as control variables.

Finally, all of the implications we have derived under the assumption that the household's time of death is certain can be derived in a model in which the time of death is uncertain. Specifically, if we assume that the husband and wife die at the same time, that the probability of the household's death in early periods is small, and that expected afterlife consumption starts in period $n + 1$ regardless of the household's time of death, then all of our results follow. Once the probability of death in a period becomes relatively large, however, we may observe religious participation declining with age. That is, once an individual is faced with a relatively high probability of death in a period it may become optimal for him to concentrate his religious participation as early as possible, since he may not survive to "invest" in future periods. Thus, the introduction of an uncertain time of death leads to the possibility of declining age–religious-participation profiles for the aged.[18]

---

[16] Formally this would require $\partial u/\partial q$ to be substantially larger than $\partial u/\partial s$ for all t. If it were not and if $\partial u/\partial s_t$ were sufficiently small relative to $\partial u/\partial s_{t-1}$, then at the optimum $r_{jt} < r_{jt-1}$. A sufficient condition for $r_{jt} > r_{jt-1}$ is that the production functions for s_t be the same in all periods and $(\partial u/\partial s_t)(\partial u/\partial s_{t-1}) \leq (1 + i)^{-1}$.

[17] Clearly the social-pressure motive for religious participation can be incorporated into the model in an analogous manner. However, except for the observation that this motive is likely to be more important in rural than urban areas, it appears difficult to specify important variables we possess which will influence it.

[18] Models in which the "date of salvation" was also a random variable were also experimented with; however it proved impossible for us to derive any implications from them.

III. Empirical Estimates

We present econometric tests of the implications of the model below using two types of data: statewide church-membership data for 1952, 1936, and 1926; and survey data on the annual frequency of church attendance for approximately 1,500 U.S. adults in 1973. We caution that the model yields implications concerning the total time devoted to religious activities, and this is unlikely to be perfectly correlated across individuals (or states) with the annual frequency of church attendance (or church membership). For example, in some religions prayer in the home is a substitute for formal church attendance. Nevertheless, these two types of data are the closest available proxies for the desired dependent variable.

A. Church-Membership Equations

Church-membership data is notoriously poor in quality. Since 1916 four censuses have been attempted, the first three (1916, 1926, 1936) by the U.S. government; the last (1952) by the National Council of the Churches of Christ in the United States of America. All suffer from reporting errors and incomplete coverage. The latter census, based upon data reported through national religious organizations, does not contain data for a number of Negro churches; thus church-membership figures are substantially underestimated in 1952 for states with relatively large black populations. The government census data was derived from questionnaires mailed to local churches. Although no explicit mention is made of systematic reporting errors in any of the census volumes, it appears that due to incomplete coverage or differential response rates church membership in rural areas and among the black population may have been underestimated in these years.

The definition of what constituted a church member also differed across religions and over time. The Catholic church considered all people who had been baptized, including infants, as members in all four years. Most Protestant sects reported only adult members (usually over age 13). For the last three sets of data, the Jewish membership figures reflected a "cultural count," the approximate number of Jews living in communities with a synagogue independent of their actual synagogue affiliation. In 1916, however, only Jewish male heads of households were reported as members. Clearly, reported membership rates varied across states simply because of variations in the religious composition of the population.

In spite of the questionable quality of the data, we have estimated cross-section regression equations for each year in an attempt to explain the interstate variations in church-membership rates. The explanatory variables which are directly related to our model include measures of the age and sex structure of the state's population, measures of male and female average wage rates, and measures of nonlabor income or wealth. To control for the measurement errors in the reported church-membership-rate variable, we also include as explanatory variables the percentages of a state's reported church members who are Catholic and Jewish, the percentage of a state's population that is nonwhite, and the state's population density.[19]

The first two rows of table 3 present estimates based upon the 1952 data with all variables entered in linear form, while the last two rows contain estimates based upon a logit transformation of the dependent variable.[20] For both specifications, as expected, the greater the percentage of the population that is female, the higher the church-membership rate. The wealth variable measured either by median family income or the median value of owner-occupied housing also has a positive effect. Both the male- and female-wage variables, as measured by the median earnings of full-year workers, have negative effects on church-membership rates. Finally, the median-age variable is positively related to church-membership rates.[21] All of these relationships are statistically significant and conform to the predictions of our model. Furthermore, each of the control variables which is statistically significant has the anticipated sign.

Similar estimates for the 1936 and 1926 data are found in columns 1–8 of table 4.[22] Unfortunately, separate wage data for males and females are unavailable for those years, and a composite wage variable—average annual earnings of manufacturing employees in the state—is used instead. Quite strikingly, we again observe significant positive wealth or (nonlabor) income coefficients and significant negative wage coefficients for both years, in accordance with the model's predictions.

In contrast with the previous results, increases in the median age of the population are associated with lower church-membership rates in 1936 and 1926, while increases in the proportion of the population over

---

[19] The latter two variables are replaced by the white proportion of the state's population and the proportion of the population residing in urban areas, for years prior to 1952. Since these variables are included to control for systematic measurement errors in the reported state membership data, we are unable to test if membership rates actually vary systematically with these variables, as our model with the consumption motive included would predict. Our wage-rate variable is actually the product of the wage rate and annual hours of work, and hence may also be subject to systematic measurement errors.

[20] We include the latter form because the linear specification does not restrict the predicted value of the dependent value to lie in the (0, 1) range. Since qualitatively the estimates are identical for both specifications, we report only the linear specification in later tables.

[21] The negative coefficient for the fraction of the population greater than age 65 may reflect either the high probabilities of death or poorer health of this group.

[22] Similar equations were estimated using the 1916 data, but significant measurement errors present in that year's independent variables lead to primarily statistically insignificant (and biased) coefficients. We thus have omitted the reporting of these results.

TABLE 3

Church-Membership Equations, 1952 (Absolute Value *t*-Statistics

Dep. Var./Ind. Var.	CATH	JEWH	NONW	DENS	MEDA	AGED	FEML	MEDY	MEDV	WAGM	WAGF	R^2
MEMB	0.204	0.040	−1.027	0.002	1.769	−5.186	2.373	. . .	0.003	−0.002	−0.041	.699
	(3.3)	(0.1)	(5.4)	(1.7)	(2.2)	(4.5)	(1.7)		(2.3)	(0.2)	(3.6)	
MEMB	0.204	0.138	−0.934	0.001	1.702	−6.393	3.878	0.024	. . .	−0.021	−0.041	.756
	(3.7)	(0.4)	(5.5)	(1.5)	(2.5)	(5.7)	(3.0)	(4.0)		(2.6)	(4.1)	
LOGM	0.085	0.006	−0.431	0.001	0.737	−2.458	1.013	. . .	0.002	−0.001	−0.017	.697
	(3.2)	(0.1)	(5.4)	(1.7)	(2.2)	(4.6)	(1.7)		(2.4)	(0.3)	(3.6)	
LOGM	0.085	0.049	−0.393	0.001	0.715	−2.705	1.651	0.010	. . .	−0.009	−0.017	.751
	(3.7)	(0.4)	(5.4)	(1.5)	(2.5)	(5.7)	(3.0)	(3.9)		(2.7)	(4.0)	

Sources.—All variables (except for the church-membership data) are reported for 1950 and calculated from *Churches and Church Membership in the U.S.* (New York: Nat. Council Churches of Christ USA, 1956); U.S. Bureau of the Census, *County and City Data Book 1952* (Washington: Government Printing Office, 1953); *1950 Census of Population: Characteristics of the Population* (Washington: Government Printing Office, 1952).

Note.—n = 49 for all equations. All LOGM coefficients have been multiplied by a factor of 10. MEMB = percentage of state population that is church members; CATH = percentage of church members that are Catholic; JEWH = percentage of church members that are Jewish; NONW = percentage of state population that is nonwhite; DENS = state population density; MEDA = median age of the state population; AGED = percentage of the state population age 65 and older; FEML = percentage of the state population that is female; MEDY = median family income in the state; MEDV = median value of owner-occupied housing in the state; WAGM = median annual earnings of male full-year workers; WAGF = median annual earnings of female full-year workers; LOGM = logarithm of (MEMB/1 — MEMB).

TABLE 4

CHURCH-MEMBERSHIP EQUATIONS, 1936 AND 1926 (Absolute Value *t*-Statistics)

	1936 EQUATIONS				1926 EQUATIONS				MEMBERSHIP-SEX RATIO EQUATIONS 1936	MEMBERSHIP-SEX RATIO EQUATIONS 1926
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
CATHYR	0.208	0.188	0.197	0.188	0.148	0.121	0.147	0.128	0.167	0.167
	(3.2)	(3.0)	(3.2)	(3.1)	(2.4)	(2.1)	(2.6)	(2.3)	(3.3)	(3.8)
JEWHYR	0.038	−0.006	0.312	0.185	−0.035	0.010	0.162	0.119	−0.044	0.100
	(0.1)	(0.0)	(1.0)	(0.6)	(0.1)	(0.0)	(0.5)	(0.4)	(0.2)	(0.5)
URBYNR	0.249	0.201	0.344	0.279	0.287	0.222	0.359	0.282	...	....
	(1.9)	(1.6)	(2.6)	(2.0)	(2.4)	(2.0)	(3.1)	(2.4)		
WHTEYR	0.297	0.429	0.032	0.213	0.175	0.371	0.006	0.220	...	...
	(2.4)	(3.3)	(2.0)	(1.0)	(1.7)	(3.2)	(0.1)	(1.4)		
SEXRYR	−0.699	0.193	0.254	0.240	−0.625	0.013	0.032	0.263	0.421	0.355
	(2.6)	(0.6)	(0.5)	(0.5)	(3.2)	(0.5)	(0.1)	(0.8)	(2.0)	(2.1)
MEDAYR	−2.837	−2.911	−5.942	−4.874	−2.896	−2.862	−5.149	−4.257	...	...
	(4.4)	(4.8)	(4.3)	(3.1)	(5.2)	(5.6)	(4.9)	(3.8)		
MEDV40	0.003	0.006	0.004	0.006	0.002	0.005	0.003	0.005	...	...
	(1.9)	(3.0)	(2.4)	(2.8)	(1.4)	(2.9)	(2.0)	(2.8)		
WAGEYR	...	−0.022	...	−0.015	...	−0.029	...	−0.022	...	...
		(2.5)		(1.4)		(2.9)		(2.0)		
AGEDYR	...	...	6.051	3.876	...	...	4.852	2.986	−0.468	−1.377
			(2.5)	(1.3)			(2.5)	(1.4)	(1.3)	(2.0)
LFPMYR	...	...	...	...	...	...	...	...	−0.858	−1.520
									(1.3)	(2.5)
LFPFYR	...	...	...	...	...	...	...	...	−0.321	−0.042
									(1.5)	(0.2)
R^2	.615	.666	.665	.680	.663	.723	.707	.734	.488	.523

SOURCES.—All variables are calculated from data in the following sources available from the Government Printing Office, Washington: U.S. Bureau of the Census, *Religious Bodies 1936* (1941); *Religious Bodies 1926* (1930); *1940 Census of Housing: General Characteristics* (1943); *1940 Census of Population: Characteristics by Age* (1943); *1940 Census of Population: Characteristics of the Population by State* (1943); *1935 Biennial Census of Manufacturers* (1938); *1925 Biennial Census of Manufacturers* (1928).

NOTE.—$N = 49$ for all equations. CATH = percentage of state church members that are Catholic; JEWH = percentage of state church members that are Jewish; URBN = percentage of state population in urbanized areas; WHTE = percentage of state population that is white; SEXR = males/100 females in the state; MEDA = median age of the population; MEDV = median value of owner occupied housing in 1940; WAGE = average yearly earnings of manufacturing production workers; AGED = percentage of the population older than age 65; LFPM = male labor force participation rate; LFPF = female labor force participation rate; YR = 1936 or 1926 value of the variable depending upon the year of the dependent variable.

age 65 lead to increases in the membership rate each year. These results are consistent with our model since the median age in both years in most states was less than 30 years, which may well lie in the downward-sloping stage of the age-participation profile. This does not explain, however, why the median-age and "aged"-population coefficients estimated from the 1952 data had the opposite sign pattern.[23]

Similar to the 1952 results, increases in the ratio of males to females is seen to decrease church membership in both 1936 and 1926 (eqs. [1] and [5]), as predicted by our model. However, the addition of either the wage variable or the percentage of the population older than 65 to the equation causes this variable to lose its significance. The average-wage variable is highly correlated with the sex-ratio variable, and this may have led to collinearity problems.[24]

Finally, in columns 9 and 10 of table 4 we present estimates of equations which seek to explain why the ratio of male/female church members varies across states. Not surprisingly, the male/female population ratio is an important determinant of the membership ratio. Moreover, in at least one year, 1926, increases in the male labor force participation rate also led to decreases in the male/female membership ratio. In contrast, the female labor force participation rate does not appear to influence the membership ratio.[25]

Overall, these results are mixed. The 1952 data, in which we have the most confidence, strongly support our model. However, the 1936 and 1926 census data, while supporting the existence of income and substitution effects on religious participation, are more ambiguous about the existence of age and sex effects. To ascertain if the latter exist, we turn to a study of variations in the frequency of church attendance across individuals.

B. Frequency-of-Church-Attendance Equations

The *Spring 1973 General Social Survey* conducted by the National Opinion Research Center contains data for a sample of 1,504 U.S. adults, including their reported annual frequencies of church or synagogue attendance. Individuals in the sample were also asked if they believed in a life after death. Approximately 76 percent of the Protestants, 70 percent of the Catholics, and 17 percent of the Jews responded affirmatively to this question (table 1B). The inclusion of these two questions makes this survey the best available source of data to empirically test our model; however it has several important weaknesses, which we note below.

---

[23] The sign reversal may be due to the high positive correlation between the two age variables in each year. As is well known, collinearity does not cause biased coefficient estimates. However, it does cause coefficient estimates to become very sensitive to the particular set of sample data used. Moreover, if two variables are highly positively correlated, the deviations of the estimated coefficients from their true values are likely to be equal and opposite in sign (see Johnson [1972, pp. 160–62] for an elaboration of these points).

[24] This correlation is a mechanical one, as an increase in the proportion of males in a state will lead to an increase in the observed average wage because males receive higher wages than females.

[25] The labor force participation rates are used as imperfect proxies for the unavailable wage-rate data by sex. Their use also leads to possible simultaneous equation bias. Unfortunately, while separate wage data by sex is available for 1952, church-membership data by sex is not.

Ideally, one would like to estimate an empirical model in which the frequency of church attendance of both the husband and wife are simultaneously determined. Unfortunately, frequency of church attendance is reported in the survey only for the individual respondents and not for their spouses. We are thus limited to explaining variations in church attendance across individuals. A sufficient number of "family-background" variables is reported by each individual, however, to enable us to specify an estimating equation that represents a reduced-form equation derived from our household model.

The second major weakness of the survey is its failure to include any information on market wage rates of the household members. Instead, information is provided on the Hodge-Siegel-Rossi occupational prestige index for the respondent and his (her) spouse, based on the census two-digit occupations in which they are employed. This index is a rating of occupational titles made by a national sample of the American population. The ratings tend to be stable over time, and for both 1960 and 1970 census data the correlation between mean annual income in an occupation and the prestige index exceeded .70.[26] As such, they may be considered extremely crude proxies for the relative market wage rates of individuals, abstracting from the individuals' positions in their life cycles. Clearly, substantial measurement errors are involved in the use of these proxies, and, making standard assumptions, one can show that their coefficients will be biased toward zero.

The survey data permit us to specify an estimating equation in which an individual's annual frequency of church attendance is related to variables predicted by the model to influence church attendance as well as a set of control variables. The former include the individual's age and sex, the occupational prestige scores for the respondent and spouse, family income, and whether the individual believes in an afterlife.[27] The control variables include the respondent's race, whether the spouse is of the same major religion, the number of children between the ages of six and 18 in the household, the number of children younger than age six in the household, and a set of dichotomous variables which indicate how the individual rates his (her) health.

---

[26] The multiple correlation using both census mean annual income and median education level in the occupation is about .86. This information, as well as the description of the Hodge-Siegel-Rossi occupational prestige index in the text, was kindly provided to us by Peter Rossi in private correspondence dated October 29, 1973.

[27] Family income was reported in 12 discrete intervals. We constructed a continuous variable by assigning to each household a family income equal to the midpoint of the interval in which the household's family income was located. In the calculation of the estimates reported in the text, we arbitrarily assigned a family income of $30,000 to the approximately 6 percent of the households whose incomes fell in the open interval of

The first three control variables are included to capture factors which influence the current utility the household receives from participating in religious activities. That is, the social value of religious activity is probably higher for racial minorities, as discrimination may limit their market consumption alternatives; for families in which both spouses are of the same religion; and for families with school-age children. In contrast, the presence of preschool-age children or an individual's being in poor health are likely to increase the cost of church attendance to the individual and thus reduce his frequency of attendance. The latter variable may also influence frequency of church attendance if it is associated, *ceteris paribus*, with higher probabilities of death. In this case, it may be optimal for the individual to concentrate his religious participation as early as possible.[28]

Table 5 presents estimates of this empirical church-attendance equation for five subsamples of the data consisting of all Protestants, Catholics, Methodists, Lutherans, and Baptists, respectively.[29] These regressions are not restricted to the set of families in which two spouses are present and earning labor income, as our theoretical model assumes. The imposition of such restrictions would reduce each subsample to approximately one quarter of its original size, leaving us only limited degrees of freedom. However, as the last column of the table indicates, imposing these restrictions for the large all-Protestant sample and reestimating the equation did lead to results which are quite similar to those found in the unrestricted sample.[30]

The estimates presented in table 5 provide, in the main, strong support for our theory. Frequency of attendance increases with age (x_1) for all groups. Females attend church more frequently than males (x_2) and nonwhites more frequently than whites (x_3). Belief in an afterlife (x_4) is quantitatively a major determinant of church attendance for all groups,

---

$25,000 or greater. Regressions were also estimated in which we assigned these families incomes of $44,000, the figure estimated by Miller (1963) to be the mean income in the open interval; however, the estimated coefficients were virtually identical with those reported in the text. Finally, we included a (1, 0) variable in the estimating equations in an attempt to control for the approximately 7 percent of the sample who failed to report their family incomes.

[28] Preliminary estimates indicated that, *ceteris paribus*, region of the country, rural-urban dichotomy, and marital status per se (independent of its influence on other variables) did not significantly influence frequency of attendance. These variables are thus not included in the regressions reported in the text.

[29] Equations were also estimated for subsamples consisting of Jews, Episcopalians, and Presbyterians. Due to the relatively small subsample sizes (table 1B), most coefficients proved to be statistically insignificant. Consequently, these results are not reported in the text.

[30] Similar results were found for the other religious groups. In each case, when a coefficient estimate remained statistically significant when the relatively small restricted sample was used, its sign was the same as that reported in table 5.

TABLE 5

FREQUENCY OF CHURCH-ATTENDANCE REGRESSIONS (Absolute Value *t*-Statistics)

Variable	Protestant	Catholic	Methodist	Lutheran	Baptist	Protestant B
x_1	0.233	0.369	0.254	0.243	0.310	0.343
	(3.4)	(3.7)	(2.0)	(1.5)	(2.4)	(1.7)
x_2	−7.241	−12.875	−7.179	−9.062	−17.373	−10.559
	(3.4)	(4.4)	(1.9)	(1.8)	(4.6)	(2.5)
x_3	−8.825	−0.255	−5.242	−26.261	−0.342	−23.412
	(3.2)	(0.1)	(1.0)	(1.1)	(0.1)	(3.9)
x_4	21.044	13.661	14.889	16.521	16.849	25.896
	(9.1)	(4.5)	(3.6)	(3.3)	(3.9)	(5.2)
x_5	6.281	11.539	7.393	−2.186	10.808	−4.141
	(2.4)	(3.5)	(1.4)	(0.5)	(2.2)	(0.6)
x_6	7.907	−1.655	17.156	−4.041	7.165	22.446
	(1.7)	(0.3)	(1.9)	(0.4)	(1.0)	(1.5)
x_7	9.853(10^{-4})	0.087(10^{-4})	2.674(10^{-3})	1.819(10^{-3})	2.304(10^{-4})	2.534(10^{-3})
	(1.8)	(0.4)	(2.6)	(1.4)	(0.7)	(1.7)
x_8	−5.506	−8.563	−1.753	−2.482	−9.114	−9.549
	(2.3)	(2.6)	(0.4)	(0.4)	(2.0)	(2.0)
x_9	−5.813	−14.911	2.413	−9.092	−11.573	−10.803
	(2.0)	(3.7)	(0.5)	(1.4)	(2.2)	(1.7)
x_{10}	−4.683	−21.654	−5.178	−0.469	−9.071	3.596
	(1.1)	(3.8)	(0.7)	(0.1)	(1.2)	(0.3)
x_{11}	0.035	0.067	−0.027	0.196	0.097	0.317
	(0.6)	(0.7)	(0.3)	(1.4)	(0.9)	(2.1)
x_{12}	−0.025	−0.081	−0.095	0.097	−0.238	0.116
	(0.4)	(1.1)	(0.9)	(0.8)	(1.9)	(0.7)
x_{13}	0.024	−1.223	−6.163	−2.122	−1.500	1.742
	(0.1)	(0.6)	(1.7)	(0.7)	(0.6)	(0.5)
x_{14}	1.822	0.845	0.932	0.552	2.989	2.836
	(2.3)	(0.8)	(0.6)	(0.3)	(2.3)	(1.7)
x_{15}	−3.507(10^{-8})	...	−8.815(10^{-8})	−8.155(10^{-8})	...	−9.749(10^{-8})
	(1.9)		(2.6)	(1.7)		(2.1)
R^2	.138	.201	.177	.241	.171	.247
N	940	388	198	124	312	221

NOTE.—Males and females treated symmetrically, i.e., if a female is being interviewed, the spouse refers to a male. Protestant B sample restricted to households in which both spouses are present and working. x_1 = age; x_2 = male = 1, female = 0; x_3 = white = 1, nonwhite = 0; x_4 = believe in afterlife, yes = 1, no or uncertain = 0; x_5 = spouse is of same religion; x_6 = no income data reported for family of (1, 0); x_7 = family income: x_8 = respondent rates health as good (1, 0); x_9 = respondent rates health as fair (1, 0); x_{10} = respondent rates health as poor (1, 0); x_{11} = Hodge-Siegel-Rossi occupational prestige score; x_{12} = spouse's occupational prestige score; x_{13} = number of children under age 6; x_{14} = number of children ages 6–18; x_{15} = family income squared.

and having a spouse of the same major religion (x_5) significantly increases religious participation for most groups.[31]

Individuals in the sample rate their health status as excellent, good, fair, or poor. Dichotomous variables were defined for the latter three categories and entered into the regressions (x_8, x_9, x_{10}). Compared with those individuals in excellent health, other individuals tend to have reduced church attendance, as expected.[32] Similarly, when their coefficients are statistically significant, an increase in the number of preschool-age children (x_{13}) present reduces church attendance, while an increase in the number of school-age children (x_{14}) present increases the frequency of attendance.

The coefficients of the individual's (x_{11}) and spouse's (x_{12}) occupational prestige indices, which serve as proxies for their market wages, are insignificant, in the main.[33] This is not surprising; substantial measurement errors are likely to be present, as we are asking the national mean annual income in a broad two-digit occupation to serve as a good estimate of an individual's actual hourly wage. Moreover, the indices take on the value of zero for those individuals (or spouses) not currently employed, thereby assigning a zero opportunity cost of time to those engaged solely in nonmarket activity.[34]

---

[31] We face a possible simultaneity problem here because frequency of church attendance may influence an individual's belief in an afterlife. A structural model would thus have attendance influencing belief as well as vice versa. The resulting reduced-form "attendance equations" would be similar to those estimated in table 5, except that "belief" would be omitted. We have estimated such equations; the results are found in table A1. Although in all cases the explanatory power of the model falls with the omission of the belief variable, the coefficient estimates of the other variables appear quite similar to those found in table 5.

[32] The primarily insignificant coefficient for the "poor-health" variable (x_{10}) may indicate that only for this group is the impact of the increased "cost" of attendance offset by a higher probability of death.

[33] The one significant coefficient is a negative coefficient on spouse's status in the Baptist equation. A plausible, but not the sole, explanation for such a result is as follows: For wives who are respondents and are not in the labor force, an increase in the spouse's wages increases the market-goods available to combine with their own time in household production, thereby increasing the value of their time spent in household production relative to time spent in religious participation. This should induce them to reduce their religious participation. This argument refers to a corner solution, while our approach in the text considers only interior solutions. For more complete discussions of how a husband's wage income influences the value of a wife's time in home production, see Gronau (1973) and Willis (1973).

[34] The only other available proxies for the opportunity costs of time are the education levels of the respondents and their spouses. Equations were estimated, substituting these variables for the occupational prestige indices. In the main, the coefficients of the other variables in the analysis were unchanged. When their coefficients were significant, the respondent's education level (two cases) had a positive impact on attendance while the spouse's education level (one case) had a negative impact. Education levels may influence nonmarket production via channels other than through their impact on wages (see Michael 1973), so these results should not be considered strong evidence in favor of (or against) our approach.

In table 5, total family income is allowed to influence annual frequency of church attendance in a nonlinear way, as both income and income squared are present. For all Protestants, Methodists, and Lutherans, we observe statistically significant backward-bending relationships between frequency of attendance and total family income. That is, church attendance first increases and then decreases with family income; the maximum frequency occurring at about $14,000, $15,000, and $11,000 for the all Protestants, Methodists, and Lutherans, respectively. On the other hand, frequency of church attendance is always positively but statistically insignificantly related to family income for Catholics and Baptists.[35]

If we interpret variations in total family income as being solely due to variations in wage rates, then both of these patterns would be consistent with our model.[36] Unfortunately, variations in total family income may be due to variations in household wage rates, hours of work, number of family members employed, or nonlabor income. The data do not permit us to separate out all of these sources of variation. Consequently, the association between total family income and church attendance should not in itself be considered strong support for our approach. However, we reemphasize here the significant relationship between church-membership rates and income and wage variables that were found using statewide membership data in the preceding section.

The estimates so far presented constrain the shape of age–religious-participation profiles to be linear and to be identical for males and females. However, a strong implication of our model is that females' hours devoted to religious activities will rise more rapidly with age than will the comparable profile for males because males' age-earnings profiles tend to be steeper than females', Moreover, for males who show sharp earnings increases during their twenties, age–religious-participation profiles may be U-shaped, first declining with age and then increasing. However, for females who have relatively flat age-earnings profiles, our theory would rule out an initial declining segment of the age–religious-participation profile.

To test for these differences in the shape of age–religious-participation profiles, the equations presented in table 5 were reestimated adding age squared as an explanatory variable and allowing the coefficients of this variable and age to vary by sex. Based upon the residual sum of squares from these equations and those found in table 5, formal *F*-statistics were calculated to test the null hypothesis that age–religious-participation profiles are linear and their slopes do not vary by sex. The results are tabulated in the first column of table 6; the null hypothesis is rejected for three of the five religion categories.

---

[35] Formally, adding the square of family income to the estimating equation does not significantly increase the explanatory power of the model for the Catholic and Baptist subsamples. The interpretation requires us to assume that the occupational prestige indices do not capture variations in market wages.

[36] Suppose that both the income effect and compensated substitution effect of a wage change may be treated as parameters. If an individual is on the backward-bending segment of his market labor supply curve *and* the income effect weighted by market hours is initially larger in absolute value than the substitution effect, we will observe a backward-bending religious-participation function, first increasing and then decreasing as the individual's wage rate increases. On the other hand, if the individual is on the upward-sloping segment of his market labor supply curve *and* the income effect weighted by market hours initially is larger than the substitution effect in absolute value, then we will observe a positive relationship between his wage and his religious participation.

The remaining columns of table 6 present the actual estimated coefficients of the age and age squared variables, by sex, for these three religion categories. Also presented is the estimated "turning age" for each religion-sex group and the estimated "overtaking age" for each religion. The former is the age at which the frequency of church attendance reaches a minimum (or maximum) for a group; the latter indicates the age until which the slope of the age–frequency-of-church-attendance profile is steeper for females than for males.[37]

Most strikingly, we observe that for all three groups the overtaking age does not occur until about age 50. That is, throughout the region in which male age-earnings profiles tend to be steeper than female age-earnings profiles, the slope of the age–church-attendance profile is steeper for females than males as our theory predicts! Moreover, for two of the three male groups, church attendance is always increasing with age, while for the third group we observe a U-shaped profile. In contrast, we note that none of the female groups exhibits an initial range of religious participation declining with age. Taken together, these results provide additional strong support for our theory.[38]

IV. Concluding Remarks

This paper represents the first systematic attempt by economists to analyze the determinants of religiosity, individuals' religious commitment and participation in church-related activities. A multiperiod utility-maximization model of household allocation of time has been developed which suggests the determinants of household religious participation and includes implications about the division of participation between

---

[37] Given the lack of statistical significance of most of the individual regression coefficients used in the calculations, these estimates are subject to substantial error and should be interpreted accordingly.

[38] Each female group's frequency of attendance reaches a peak around age 60, which is not directly predicted by a model that includes only the salvation motive. The declining profile may reflect the impact of the consumption motive or that age may be a better (inverse) measure of health status than the crude self-evaluation health question for older females. The failure for the males' profiles to reach a peak is not a statistical artifact caused by the use of a quadratic function, as the use of a cubic function in age did not increase the explanatory power of the model. Rather, after retirement the consumption value of religion may increase for males. We have no firm evidence on this point, however.

TABLE 6

ESTIMATED IMPACT OF AGE AND AGE² ON FREQUENCY OF CHURCH ATTENDANCE, BY SEX

RELIGION	$F(m, n)$	FEMALE			MALE			OVERTAKING AGE
		Age	Age²	Turning Age	Age	Age²	Turning Age	
Protestant	6.923(3,922)*R*	1.548*	−.013	60 max	.052	.002	Always increasing	50
Catholic	1.843(3,370)*N*	. . .	. . .	. . .	. . .	. . .	. . .	. . .
Baptist	4.381(3,294)*R*	1.980*	−.017	58 max	.007	.002	Always increasing	52
Lutheran.	0.728(3,106)*N*	. . .	. . .	. . .	. . .	. . .	. . .	. . .
Methodist	3.620(3,180)*R*	1.473	−.012	61 max	−.623	.003	39 min	52

NOTE.—$F(m, n)$ = F-statistic to test the null hypothesis (H_0) and (degrees of freedom numerator, degrees of freedom denominator); $R(N)$ = reject (not reject) H_0 at .95 level of significance; H_0 = the age–frequency-of-attendance relationship is linear and does not vary by sex; H_a = the frequency of church attendance is a function of age *and* age squared, *and* the relationship varies by sex (alternate hypothesis); turning age = age at which frequency of church attendance reaches a minimum (min) or a maximum (max); overtaking age = age up until which the slope of the age-church attendance profile is steeper for females than males.

* Individual coefficient statistically significantly different from zero at .95 level of significance, two-tailed test.

husband and wife and the shape of a household's life-cycle religious-participation profile. Virtually all of the simple correlations of variables with religiosity, previously uncovered by sociologists, are consistent with our framework.

The theory has been tested using multivariate methods on two different types of data: statewide data on church-membership rates and survey data on the annual frequency of church attendance for 1,500 U.S. adults. Neither source of data is completely adequate for an empirical test of the model. The former suffers from reporting errors and incomplete coverage; the latter contains no wage-rate data and reports the frequency of church attendance for only one spouse. However, the empirical results based upon the two samples taken together provide strong support for our model.

The model which we have proposed in this paper can be easily extended in several directions. These extensions allow one to analyze a wide range of phenomena relating to the interaction of household allocation of time and religiosity.[39] For example, one extension is to generalize the expected afterlife-benefit-production function to allow the substitution of market goods for household time. These goods might be interpreted as being church dues, contributions to religious charities, or contributions in general. Many implications follow directly from this extension.

First, as real wages grow over time we should expect individuals to shift toward less time-intensive forms of religious activities. This could occur by the religious institutions themselves changing; such as the development of Reform Judaism, which is less time-intensive than other forms of Judaism.[40] Alternatively, if a set of religions which have relatively similar basic beliefs exists (i.e., many of the Protestant denominations), we should expect the less time-intensive religions to grow secularly relative to the more time-intensive ones. In contrast, when the unemployment rate rises, the opportunity cost of time falls for many individuals, hence we should expect to observe the reverse pattern, with more time-intensive religions growing relative to less time-intensive ones.[41] Although historical church-membership data is notoriously poor in quality, these implications could be tested in principle by using aggregate historical data on church membership by denomination.

Second, this model predicts that, if an individual faces an upward-sloping age-earnings profile, he should decrease the time intensity of his religious activities over the course of his life cycle. This could be accomplished either through interdenominational mobility toward less time-intensive denominations or through the substitution of market goods for household time within a given denomination. With respect to the former, survey data indicate that approximately 20 percent of U.S. adult males change their religious preferences at least once during their lifetimes.[42] If measures of the time-intensity of various denominations can be developed, survey data can be used to test the implications concerning interdenominational mobility. With respect to the latter, if we interpret market goods used in religious activities as contributions to churches (or donations in general), the model provides us with a theory of life-cycle contributions. Any attempts to test such a theory, however, would require elaborating it to consider factors such as the federal tax treatment of contributions.[43]

---

[39] The formal models underlying the discussion that follows are contained in an appendix that is available from the authors. In future reseach we hope to empirically test many of the hypotheses that they generate.

[40] This point has been previously suggested by Gary Becker and is cited in Linder (1970). Linder also discusses examples of similar types of "institutional change" occurring in Sweden.

[41] Some evidence on this point is offered in Sales (1972), although his analysis focuses on authoritarian versus nonauthoritarian denominations. Casual empiricism suggests that this dichotomy may also reflect differing time intensities.

Our original model as well as the generalization discussed above assume that individuals' life-cycle wage profiles are exogenous. A second extension of the model is to make life-cycle human-capital-investment decisions, and consequently wage profiles, endogenous. Under certain assumptions, one can show that the steepness of age-earnings profiles implied by such a model will vary inversely with individuals' extent of belief in an afterlife. That is, individuals who assign higher probabilities to the existence of an afterlife will tend to have relatively flatter age-earnings profiles.[44] Such a model may provide an explanation of why life-cycle occupational mobility varies across religions.[45] Survey data may also permit one to test whether a relationship exists between the extent of religious belief and the shape of age-earnings profiles across individuals of the same religion.[46]

These generalizations of the model should make it clear that we have

---

[42] See Warren 1970.

[43] Dickinson (1970) has constructed estimates of contributions to each of the major religions in the United States for the 1929–59 period.

[44] Formal details of such a model are also contained in the available appendix. Each household is assumed to have a subjective probability, Θ, that there is an afterlife and to maximize a utility function of the form

$$U^{*} = U^{1}(C_1, C_2, \dots C_n) + \Theta U^{2}(q), \qquad (1'')$$

where the expected afterlife-production function (3) is now interpreted as being conditional on belief in an afterlife. Specific assumptions about the shapes of the various functions then lead to the results cited in the text. This conclusion is very sensitive to the assumptions which we have made and is not a general result. Rather, we offer it only as a tentative hypothesis to be subject to future empirical testing.

[45] Sociologists have investigated this relationship (see Lenski 1961, Gockel 1969, Jackson, Fox, and Crocket 1970, Warren 1970, Featherman 1971, and Schuman 1971, for examples. Recently several economists, including Hause (1972) and Taubman (1974), have also observed earnings varying, *ceteris paribus*, across religions.

[46] Unfortunately, the NORC data used in this paper contain no information on individual's wage rates or annual labor earnings.

only begun the process of development and testing of a unified theory of household allocation of time and religiosity. They also suggest the directions which future empirical research on the subject should take. Nevertheless, we believe that the empirical results presented in the paper provide strong initial support for our approach and suggest that religious participation should be added to the growing set of nonmarket activities which have proven amenable to being analyzed in a household-allocation-of-time framework.

Appendix

TABLE A1

Frequency-of-Church-Attendance Regressions: Belief-in-Afterlife Variable Omitted (Absolute Value t-Statistics)

Variable	Protestant	Catholic	Methodist	Lutheran	Baptist
x_1	0.245	0.392	0.286	0.166	0.342
	(3.4)	(3.9)	(2.2)	(1.0)	(2.0)
x_2	−8.937	−13.572	−8.892	−10.531	−18.670
	(4.0)	(4.6)	(2.3)	(2.0)	(4.8)
x_3	−9.019	4.516	−5.212	−29.416	0.544
	(3.1)	(0.7)	(1.0)	(1.1)	(0.1)
x_5	8.730	10.530	7.843	−1.739	14.267
	(3.2)	(3.1)	(1.4)	(0.3)	(2.8)
x_6	5.656	−2.370	13.446	−9.297	6.690
	(1.1)	(0.3)	(1.5)	(0.9)	(0.9)
x_7	$8.651(10^{-4})$	$1.395(10^{-4})$	$2.162(10^{-3})$	$2.001(10^{-3})$	$2.192(10^{-4})$
	(1.5)	(0.6)	(2.0)	(1.4)	(0.7)
x_8	−5.758	−8.995	−1.382	−0.544	−11.742
	(2.3)	(2.7)	(0.3)	(0.1)	(2.6)
x_9	−5.649	−15.573	1.671	−8.365	−14.012
	(1.9)	(3.7)	(0.3)	(1.2)	(2.7)
x_{10}	−4.264	−23.364	−4.144	4.176	−11.537
	(1.0)	(4.0)	(0.5)	(0.4)	(1.5)
x_{11}	0.084	0.087	0.038	0.295	0.147
	(1.3)	(1.0)	(0.4)	(2.0)	(1.3)
x_{12}	−0.035	−0.058	−0.083	0.146	−0.288
	(0.6)	(0.8)	(0.7)	(1.1)	(2.2)
x_{13}	−0.348	−1.144	−6.607	−4.170	−1.965
	(0.2)	(0.6)	(1.7)	(1.2)	(0.8)
x_{14}	1.769	0.616	0.965	0.256	3.001
	(2.1)	(0.6)	(0.6)	(0.1)	(2.3)
x_{15}	$-3.371(10^{-8})$	...	$-7.274(10^{-8})$	−9.321	...
	(1.8)		(2.1)	(1.8)	
R^2	.061	.158	.118	.165	.128
N	940	388	198	124	311

Note.—See table 5 for variable definitions.

References

Argyle, M. *Religious Behavior.* Glencoe, Ill.: Free Press, 1959.

Becker, G. "A Theory of Allocation of Time." *Econ. J.* 75 (September 1965): 493–517.

———. "A Theory of Marriage: Part I." *J.P.E.* 81, no. 4 (July/August 1973): 813–47.

Bultena, L. "Church Membership and Church Attendance in Madison, Wisconsin." *American Soc. Rev.* 14 (May 1948): 384–89.

Croog, S. H., and Levine S. "Religious Identity and Response to Serious Illness: A Report on Heart Patients." *Soc. Sci. and Medicine* 6 (February 1972): 17–32.

Demerath, N. J. *Social Class in American Protestantism.* Glencoe, Ill.: Free Press, 1965.

Dickinson, F. *The Changing Position of Philanthropy in the American Economy.* New York: Columbia Univ. Press, 1970.

Featherman, D. "The Socioeconomic Achievement of White-Religio-Ethic Subgroups: Social and Psychological Explanations." *American Soc. Rev.* 36 (April 1972): 207–22.

Glock, G., and Stark, R. *Religion and Society in Tension.* Chicago: Rand McNally, 1965.

Gockel, G. "Income and Religious Affiliation: A Regression Analysis." *American J. Soc.* 74 (May 1969): 632–47.

Gronau, R. "The Intrafamily Allocation of Time: The Value of Housewives' Time." *A.E.R.* 63 (September 1973): 634–52.

Grossman, M. "On the Concept of Health Capital and the Demand for Health." *J.P.E.* 80, no. 2 (March/April 1972): 223–55.

Hamermesh, D., and Soss, N. "An Economic Theory of Suicide." *J.P.E.* 82, no. 1 (January/February 1974): 83–98.

Hause, J. "Earnings Profile: Ability and Schooling." *J.P.E.* 80, no. 3 (suppl.; May/June 1972): 108–39.

Jackson, E.; Fox, W.; and Crocket, H. "Religion and Occupational Achievement." *American Soc. Rev.* 35 (February 1970): 48–63.

Johnston, J. *Econometric Methods.* 2d ed. New York: McGraw-Hill, 1972.

Lazarwitz, B. "Some Factors Associated with Variations in Church Attendance." *Soc. Forces* 39 (May 1961): 300–309.

Leibowitz, A. "Home Investment in Children." *J.P.E.* 82, no. 2 (suppl.; March/April 1974): 111–32.

Lenski, G. "Social Correlates of Religious Interest." *American Soc. Rev.* 18 (October 1953): 533–44.

———. *The Religious Factor.* Garden City, N.J.: Doubleday, 1961.

———. "The Sociology of Religion in the United States." *Soc. Compass* 9 (1962): 307–34.

Linder, S. *The Harried Leisure Class.* New York: Columbia Univ. Press, 1970.

Michael, R. "Education in Nonmarket Production." *J.P.E.* 81, no. 2 (March/April, 1973): 306–28.

Miller, H. *Trends in the Income of Families and Persons in the United States 1947–60.* Washington: Government Printing Office, 1963.

Orbach, H. "Aging and Religion." *Geriatrics* 16 (1961): 530–40.

Roof, W. C. "The Local-Cosmopolitan Orientation and Traditional Religious Commitment." *Soc. Analysis* 33 (Spring 1972): 1–15.

———. "Explaining Traditional Religion in Contemporary Society." In *Changing Perspectives in the Scientific Study of Religion,* edited by A. Eister. New York: Wiley, 1973.

Sales, S. "Economic Threat as a Determinant of Conversion Rates in Authoritarian and Nonauthoritarian Churches." *J. Personality and Soc. Psychology* 23 (1972): 420–48.

Schuman, M. "The Religious Factor in Detroit: Review, Replication, and Reanalysis." *American Soc. Rev.* 36 (February 1971): 30–48.

Sklare, M. *Conservative Judaism: An American Religious Movement.* New York: Schocken, 1972.

Taubman, P. "Schooling, Ability, Nonpecuniary Rewards, Socioeconomic Background and the Lifetime Distribution of Earnings." Paper presented to the NBER conference on "Research in Income and Wealth," Ann Arbor, Mich., May 1974.

Warren, B. "Socioeconomic Achievement and Religion: The American Case." *Soc. Inquiry* 40 (Spring 1970): 130–55.

Willis, R. "A New Economic Approach to the Economic Theory of Fertility Behavior." *J.P.E.* 81, no. 2 (suppl.; March/April 1973): 4–65.

Substitution among Capital, Labor, and Natural Resource Products in American Manufacturing

---

David Burras Humphrey

U.S. Department of Commerce

J. R. Moroney

Tulane University

This paper presents estimates of partial elasticities of substitution among reproducible capital, labor, and an input aggregate of natural resource products. We are specifically interested in two hypotheses: (i) Are natural resource products strictly complementary in production with either capital or labor? (ii) Are resource products typically less substitutable with capital than with labor? To both questions the answer is, generally, no. Two modes of investigation are used, one based on a translog production function and the other making use of a translog cost function. For most industry groups, the estimated substitution elasticities obtained from the cost function are somewhat lower than those based on the translog production function.

I. Introduction

There are presently two conflicting views concerning the relations among reproducible capital, labor, and natural resources as productive inputs. The first, which was adumbrated by Leontief (1953, 1956) and reiterated by Vanek (1959, 1963), Weiser (1968), and Baldwin (1971), is that natural resources and reproducible capital tend to be complementary. They suggest that manufactured commodities intensive in the use of nat-

---

This is a revised version of a paper presented December 30, 1973, at the Econometric Society in New York City. An anonymous referee and Prof. Ernst R. Berndt of the University of British Columbia offered several useful suggestions, which were incorporated here. We are pleased to acknowledge the programming assistance of Messrs. Donald Kemp and Alden Toevs. The views expressed in this paper are not necessarily the official views of the Commerce Department.

[*Journal of Political Economy*, 1975, vol. 83, no. 1]

© 1975 by The University of Chicago. All rights reserved.

ural resources necessarily require the intensive use of capital (that is, resource-intensive goods have higher capital-output ratios, on the average, than non-resource-intensive commodities). These authors provide some fragmentary and usually indirect empirical support for this assertion.[1]

A contrasting view is that capital equipment, labor, and natural resources are significantly substitutable as productive inputs. Barnett and Morse (1963), for example, describe several processes whereby capital and labor may hypothetically be substituted for natural resources within a given state of technical knowledge.[2] In technologically progressive sectors, Hagen (1953), Adler (1961), Barnett and Morse (1963), and others emphasize that capital–natural resource substitution often accompanies improvements in technical knowledge and, indeed, that innovation is often directed toward increasing such substitutability. A somewhat broader framework for substitution possibilities is offered by Meier (1961), who suggests that the substitutability of one resource for another is augmented by expansion of the stock of technical knowledge or the flow of communications. In his view, increases in the stock of information in society should facilitate substitutability of both capital and labor for tangible natural resources.

The technical substitutability of capital and labor for natural resources is an interesting issue in itself. It is also important. If nonreplenishable or slowly growing natural resources must be utilized in fixed or almost fixed proportions with capital, labor, or both, the growth of the real national product in the United States and other resource-scarce economies may become severely limited in the years ahead.[3] The forecasts of economic doomsday within 30 years for the United States, recently popularized in the writings of Park (1968), Behrens (1971), and Meadows et al. (1972), are based on the hypothesis of fixed proportions among natural resources, capital, and labor (together with some stringent assumptions concerning the growth and composition of aggregate demand). Whether this is or is not a valid hypothesis, the rational design of public and private policies concerning foreign trade and the allocation of revenues for resource development require some knowledge of technical substitution possibilities.[4]

---

[1] More specifically, both Leontief (1953, 1956) (using 1947 U.S. input-output statistics) and Baldwin (1971) (using 1958 U.S. input-output data) found that the direct plus indirect capital requirements per dollar value of aggregate U.S. exports are smaller than the corresponding requirements per dollar value of aggregate U.S. import substitutes. If natural resource sectors are excluded from the calculations, however, the capital requirements of U.S. exports typically exceed those of U.S. imports (that is, the well-known Leontief paradox is dispelled). Similarly, Weiser (1968) found the labor input requirement per unit of aggregate U.S. exports to be greater than those per unit of aggregate import substitute; but the differential is smaller if natural resource sectors are excluded from the computation. Finally, Vanek (1963) discovered from the 1947 U.S. input-output tables that the direct plus indirect capital-output ratios were somewhat higher, on the average, among natural resource sectors than among non-resource-intensive sectors.

[2] The essence of their model is a "resource conversion function." According to this model, resources of continuously declining quality (where quality encompasses the dimensions of "richness" or "yield" as well as physical accessibility) are transformed into resources of "standard" quality by the more intensive use of capital and labor (see Barnett and Morse 1963, esp. pp. 109–25, 127–37). For a discussion of the shifts occurring in the resource conversion function as a consequence of technological change, see pp. 159–63. Some explicit examples are reviewed in pp. 197–99.

[3] The severity of the problem should not be overstated. Domestic scarcities of certain natural resources may be relieved by foreign trade and new resource discoveries. Reliance on foreign trade for essential resources, of course, reduces national autonomy and may be quite costly over short periods, as Western nations are acutely aware. Furthermore,

A review of the literature indicates that specific information respecting substitution among capital, labor, and natural resource products is not yet available. Indeed, no systematic study along these lines has to our knowledge ever been attempted. Accordingly, our object is to investigate these relations using, first, an explicit, three-input production function and, second, a model of constrained cost minimization. Our units for analyzing input substitution are U.S. four-digit manufacturing industries, classified by two-digit product group, for the year 1963. We initially make use of a transcendental logarithmic production function which was suggested by Christensen, Jorgenson, and Lau (1970). It represents a useful generalization by comparison with the Cobb-Douglas and the ordinary constant elasticity of substitution (CES) production functions in that one may analyze and estimate under fairly general conditions the partial elasticities of substitution among all pairs of the $n > 2$ inputs. The Cobb-Douglas form, of course, restricts all such elasticities to unity. And the multifactor CES form is undesirably restrictive either because of the need to specify a priori certain input separability conditions (Sato 1967) or because of the a priori requirement that substitution elasticities stand in fixed ratios to one another (Mukerji 1963). The cost function, which we use as an alternative framework for investigation, also has the transcendental logarithmic form and is described in Section V.

A discussion of the measurement of natural resource inputs seems appropriate. We discard at the outset my attempt to estimate flows of basic natural resources into manufacturing industries. Instead, we estimate natural resource inputs used by each industry according to quantity indexes of natural resource products purchased. (Natural resource products are themselves the first-stage results in the physical transformation of

---

the law of conservation of energy suggests that no natural resources are strictly nonrenewable, since they may hypothetically be reembodied in a variety of productive forms. The extent to which this law applies empirically is diminished by the second law of thermodynamics, which continuously increases entropy in any economic system (see Georgescu-Roegen 1971).

[4] We should perhaps reiterate here that our focus in this paper is restricted to technical substitution among inputs within given classes of products. Other types of substitution designed to relieve the increased scarcity of domestic resources, such as product substitution in final demand or importation, are beyond the scope of this paper.

pristine natural resources.) We do so for two reasons. The first is a practical limitation in available statistics. Accurate industrial usage of basic natural resources simply does not exist, but detailed usage of natural resource products by four-digit manufacturing industries may readily be estimated. (For a detailed description of these computations, see Appendix A.) The second reason is conceptual. We believe that the economic analysis of natural resource products as inputs is more enlightening than an analysis of natural resources, even if the latter were feasible. The reason is that pure natural resources are an ill-defined collection whose elements depend in an essential way on mankind's technical knowledge. For example, pitchblende and uraninite were not regarded as natural resources earlier in this century because of limited knowledge of fission technology. And oil-rich shale has only recently become recognized as a potentially important, economically exploitable natural resource in this country. Thus, the definition of basic natural resources depends on technology and prices and is constantly changing in technologically progressive economies (see Schultz 1953; Potter and Christy 1962, p. 2).

Accordingly, we construct for each manufacturing industry an input aggregate of 19 natural resource products used in production (see Appendix B). We consider the degree to which capital and labor may be substituted in production for this resource product aggregate, which is regarded analytically as a third explicit factor of production. These 19 resource product categories are essentially the same ones used by other economists concerned with the influence of natural resources on the composition of foreign trade (Leontief 1956; Vanek 1963; Weiser 1968; Baldwin 1971) and the role of resources in economic growth (Barnett and Morse 1963, esp. chap. 8).

We assume that the input aggregate of resource products enters production during the year in which it is purchased. We thus make no allowance for changes in inventory holdings of resource products. Because 1963 was a year of comparatively smooth economic expansion without rapid changes in aggregate business inventories, this assumption would seem to be reasonable.

Section II briefly describes the transcendental logarithmic production function (denoted hereafter as "translog") and some of its salient properties. In Section III the statistical model and measurement of variables are discussed. Section IV presents the production function estimates, together with estimates of the partial elasticities of substitution. We proceed to test two hypotheses: (1) that capital-resource substitutability differs from labor-resource substitutability; and (2) that production functions within the natural resource product-intensive sectors differ from those within sectors making exiguous use of resource products. Section V presents an alternative framework for investigating input substitution, based on a cost function instead of a production function, together with some statistical results. Section VI summarizes and concludes the paper.

II. The Production Function Framework

A translog production function describing the relation between physical output and input services from three productive factors may be written

$$\ln q = \ln \alpha_0 + \sum_i \alpha_i \ln X_i + \tfrac{1}{2} \sum_i \sum_j \gamma_{ij} \ln X_i \ln X_j, \qquad (1)$$

where q denotes output, α_0 represents the state of technical knowledge, and X_i and X_j are services of inputs. Here α_i and γ_{ij} are technologically determined parameters. Equality of γ_{ij} and γ_{ji} $(i \neq j)$ is necessary for the applicability of Young's theorem to integrable functions, although such equality could in principle be tested (see Berndt and Christensen 1973*b*).

If one disregards the log-quadratic terms, (1) is simply a three-input Cobb-Douglas function in which α_i is the strictly positive output elasticity of the *i*th input. Yet if one or more γ_{ij} is non-zero, (1) is distinctly different from the Cobb-Douglas form. Note also that all input levels must be strictly positive, since if $\ln X_i$ or $\ln X_j \to -\infty$, output becomes ill defined. If one adopts the maintained hypothesis that (1) is subject to constant returns to scale, the following restrictions pertain:

$$\sum_i \alpha_i = 1, \sum_i \gamma_{ij} = \sum_j \gamma_{ji} = \sum_i \sum_j \gamma_{ij} = 0, \qquad \text{for } i, j = K, L, N.$$

One may denote the marginal product of labor as

$$f_L = \frac{q}{X_L} (\alpha_L + \sum_j \gamma_{Lj} \ln X_j), \qquad j = K, L, N. \qquad (2)$$

From (2) it is clear that, given the usage of a finite amount of X_L, f_L becomes negative either as $X_j \to 0$ and $\gamma_{Lj} > 0$ or as X_j increases indefinitely and $\gamma_{Lj} < 0$ $(j = K, L, N)$. Thus, the translog function possesses uneconomic regions over certain ranges of input space.

The direct second-order derivative and the cross partial derivative with respect to capital are

$$f_{LL} = \frac{q}{X_L^2} [\gamma_{LL} + (\alpha_L + \sum_j \gamma_{Lj} \ln X_j - 1)(\alpha_L + \sum_j \gamma_{Lj} \ln X_j)], \qquad j = K, L, N, \qquad (3)$$

$$f_{KL} = \frac{q}{X_K X_L} [\gamma_{KL} + (\alpha_K + \sum_j \gamma_{Kj} \ln X_j)(\alpha_L + \sum_j \gamma_{Lj} \ln X_j)], \qquad j = K, L, N. \qquad (4)$$

Using the analogues of equations (2), (3), and (4) for all inputs, one may estimate the values of the first and second derivatives from a knowledge of the production function parameters and the levels of input services.

The economic region of a linearly homogeneous production function is characterized by strictly positive marginal productivities of all inputs. A convenient way to express this with reference to (1) is

$$\frac{\partial q}{\partial X_i}\frac{X_i}{q} = \frac{\partial \ln q}{\partial \ln X_i} = \alpha_i + \sum_j \gamma_{ij} \ln X_j > 0, \qquad i, j = K, L, N. \tag{5}$$

Thus, all output elasticities must be positive. These elasticities are in general not constant but depend on the levels of input services. Assuming that input and product markets are competitive, the necessary conditions for profit maximization are

$$f_i = p_i, \qquad i = K, L, N, \tag{6}$$

where p_i is the factor price of the *i*th input relative to the price of output (Second-order conditions are discussed with reference to eq. [11a].) Substituting (6) into (5) yields the necessary conditions for economic efficiency with reference to distributive shares,

$$M_i = \frac{p_i X_i}{q} > 0, \qquad i = K, L, N, \tag{7}$$

where M_i is the relative share of the *i*th input in total cost. Constant returns to scale together with purely competitive markets assure that the sum of these shares exhausts total cost.

Following the rigorous demonstration by Berndt and Christensen (1973*a*), we offer an economically intuitive explanation of the correspondence between weak functional separability of inputs and the equality of certain partial elasticities of substitution (as defined according to Allen [1938]). The marginal rate of technical substitution between inputs *K* and *L*, for example, is defined by

$$R_{KL} = \frac{f_L}{f_K}. \tag{8}$$

Leontief (1947, p. 364) showed that the mathematical condition for two inputs, *K* and *L*, to be (weakly) functionally separable from a third input *N* is that the ratio (f_L/f_K) be independent of the level of *N*. This mathematical result may be explained intuitively. Suppose the usage of *K* and *L* is held constant and the usage of *N* increases. If the augmented flow of *N* renders *K* and *L* economically more effective at the margin and raises their individual effectiveness by exactly the same relative amount, *K* and *L* are functionally separable from *N*. In other words, the greater usage of *N* shifts the marginal product schedules of *K* and *L* proportionately (which is observationally the same as a familiar Hicks-neutral technological change). Positing that *N* is an explicit third input, it is intuitively reasonable that, if a change in its usage shifts the marginal product schedules of K and L vertically in the same proportion, N bears an equally close substitution or complementary relationship to both inputs: the partial elasticity of substitution between K and N is the same as the partial elasticity of substitution between L and N.[5]

If (8) is differentiated with respect to N, the mathematical condition for functional separability of K and L from N is

$$f_K f_{LN} - f_L f_{KN} = 0. \tag{9}$$

Obtaining these derivatives from (1) and substituting into (9) yield the condition for functional separability in the translog function:

$$M_K \gamma_{LN} - M_L \gamma_{KN} = 0. \tag{10}$$

Accordingly, if the weighted difference between the parameters γ_{LN} and γ_{KN} is zero, one may infer that the partial elasticities of substitution σ_{KN} and σ_{LN} are equal. Note that the values of M_K and M_L generally depend on input usage; thus, equation (10) entails a test of $\sigma_{KN} = \sigma_{LN}$ at a point in input space. Equivalently, equation (10) is a test for point separability, not global separability. A stochastic version of (10) is the basis for our empirical tests in Section IV.

Having discussed functional separability, we should now state that the following analysis relies on the assumption that inputs K, L, and N are weakly separable from nonresource intermediate inputs, I. That is, we assume that the production function in each industry may be written

$$V = F(K, L, N, I) = J[f(K, L, N), I],$$

where V is the gross output attributable to the inputs K, L, N, and I and $q = f(K, L, N)$ is the measure of net output in (1).

The econometric implication is that the estimated production and cost parameters are assumed to be invariant to the usage of nonresource intermediate inputs. This is a strong assumption, and one's interpretation of the statistical estimates hinges on its applicability. The issue is sufficiently important to merit a brief discussion. It is well known that the outputs of certain industries, both within and outside the manufacturing sector, are closely linked with those of other industries within manufacturing. The essence of our separability assumption is this: the process of netting out the value of all such nonresource intermediate goods is assumed to leave the relative shares of K, L, and N in the value of gross output, V, proportional to their corresponding shares in the value of net output, q.

A sufficient (but not a necessary) condition ensuring this result is that the relative cost shares of the aggregate of nonresource intermediate inputs be constant across all four-digit industries within a two-digit product group. They are not. But a less stringent condition pertaining to relative shares is consistent with separability. Let the shares of K, L, N, and I in the values of V for industries a and b be $\{S_{K(a)} = .2, S_{L(a)} = .4, S_{N(a)} = .2, S_{I(a)} = .2\}$; $\{S_{K(b)} = .2, S_{L(b)} = .2, S_{N(b)} = .1, S_{I(b)} = .5\}$. If, under conditions of static technology and fixed input prices, the relative shares in the values of net output, q, were $\{M_{K(a)} = .25, M_{L(a)} = .50, M_{N(a)} = .25\}$; $\{M_{K(b)} = .40, M_{L(b)} = .40, M_{N(b)} = .20\}$, the separability assumption would be satisfied. One should ideally wish to test this hypothesis. It is not feasible to do so, however, because price and quantity indexes for nonresource intermediate inputs are not available. The construction of such indexes would itself be a large-scale project requiring probably several years of work. Some idea of the scope of such an undertaking may be found in Carter (1970).

---

[5] Berndt and Christensen (1973*a*) provide a complete and rigorous demonstration of the conditions necessary for weak and strong input separability.

Allen (1938, p. 504) defined the partial elasticity of substitution between inputs i and j by

$$\sigma_{ij} = \left(\sum_{i} f_i X_i / X_i X_j\right) (|F_{ij}| / |F|), \qquad i, j = K, L, N, \tag{11}$$

where

$$|F| = \begin{vmatrix} 0 & f_K & f_L & f_N \\ f_K & f_{KK} & f_{KL} & f_{KN} \\ f_L & f_{LK} & f_{LL} & f_{LN} \\ f_N & f_{NK} & f_{NL} & f_{NN} \end{vmatrix} \tag{11a}$$

and $|F_{ij}|$ is the cofactor of f_{ij} in $|F|$. The assumption of linear homogeneity of (1) assures that $\sum_i f_i X_i \equiv q$. As equations (2) to (4) indicate, the values of the elements of $|F|$, and therefore the numerical values of σ_{ij}, generally vary with the levels of input usage. In Sections IV and V we therefore present point estimates of the partial elasticities evaluated at the sample means. Finally, it is well known that a sufficient condition for stability of employment equilibrium is that the elements of $|F|$ be associated with a negative definite quadratic form. In the three-input model, the bordered principal minors of $|F|$ must be positive and negative, respectively (see Allen 1938, p. 503).

III. Statistical Considerations

A. The Regression Model

We estimate the parameters of (1) using stochastic versions of equations (5). Substituting the distributive share of each input for its output elasticity, these equations are

$$M_L = \alpha_L + \gamma_{LL} \ln L + \gamma_{LK} \ln K + \gamma_{LN} \ln N + u_L, \tag{12a}$$

$$M_K = \alpha_K + \gamma_{KL} \ln L + \gamma_{KK} \ln K + \gamma_{KN} \ln N + u_K, \tag{12b}$$

$$M_N = \alpha_N + \gamma_{NL} \ln L + \gamma_{NK} \ln K + \gamma_{NN} \ln N + u_N. \tag{12c}$$

The equality of relative shares and corresponding output elasticities is established by profit-maximizing behavior in competitive markets. Hence, the disturbances in (12a)–(12c) may be attributed to a variety of forces, including empirical deviations from purely competitive markets and the inability of entrepreneurs instantaneously to maximize profit.[6]

The assumption of linear homogeneity of (1), together with the symmetry restrictions, ensures that parameter estimates of any two of equations (12) will identify exactly all parameters of the production function. Consider, for example, the estimation of (12a) and (12b). The restriction that $\alpha_L + \alpha_K + \alpha_N = 1$ ensures that $\hat{\alpha}_N = 1 - \hat{\alpha}_L - \hat{\alpha}_K$. Similarly, the symmetry restriction ensures that $\hat{\gamma}_{ij} = \hat{\gamma}_{ji}$ $(i \neq j)$. Finally, the restriction of zero row sums of the γ's requires that

$$\hat{\gamma}_{ij} = -(\hat{\gamma}_{ii} + \hat{\gamma}_{ik}) \qquad \text{for } i, j, k = K, L, N.^7$$

Thus, only eight of the 12 parameters in (12a)–(12c) are unrestricted.

Two distinct estimation problems must be solved. First, one cannot estimate a set of unique parameters for any two equations of the system (12a)–(12c) by applying least squares to each equation individually, since the estimates $\hat{\gamma}_{ij}$ and $\hat{\gamma}_{ji}$ $(i \neq j)$ will not generally be equal. Imposing the symmetry and linear homogeneity restrictions allows one to write (12a) and (12b) in alternative forms suitable for estimation:

$$M_L = \alpha_L + \gamma_{LL} \ln (L/N) + \gamma_{LK} \ln (K/N) + u_L, \tag{13a}$$

$$M_K = \alpha_K + \gamma_{KL} \ln (L/N) + \gamma_{KK} \ln (K/N) + u_K. \tag{13b}$$

Equality of the estimators $\hat{\gamma}_{LK}$ and $\hat{\gamma}_{KL}$ may be imposed by combining (13a) and (13b) in a single regression equation:

$$\begin{bmatrix} M_L \\ M_K \end{bmatrix} = \begin{bmatrix} 1 & 0 & \ln (L/N) & \ln (K/N) & 0 \\ 0 & 1 & 0 & \ln (L/N) & \ln (K/N) \end{bmatrix} \begin{bmatrix} \alpha_L \\ \alpha_K \\ \gamma_{LL} \\ \gamma_{LK} \\ \gamma_{KK} \end{bmatrix} + \begin{bmatrix} u_L \\ u_K \end{bmatrix}. \tag{14}$$

The disturbances u_L and u_K are likely to be correlated, because random deviations from profit maximization should affect both the markets for labor and capital. Hence, the two-stage estimation procedure suggested by Zellner (1962) may be expected to yield more efficient parameter estimates.

The second problem is computational. The estimates obtained by applying ordinary least squares or Zellner two-stage least squares depend on which two equations from (12a)–(12c) are chosen. Maximum-likelihood estimates would, of course, be independent of which equations were selected. Kmenta and Gilbert (1968) showed in a series of Monte Carlo experiments that if one iterates the Zellner two-stage procedure, the iterative Zellner (IZEF) estimates and maximum-likelihood estimates are identical. Ruble (1968) showed formally that the IZEF and maximum-likelihood methods are computationally equivalent. Accordingly, we estimate parameters for each industry group by applying the IZEF method to equation (14). These estimates are asymptotic maximum-likelihood estimates and are therefore independent of our using the labor and capital share equations.

---

[6] Cf. Mundlak and Hoch (1965) and Zellner, Kmenta, and Dreze (1966) for detailed analyses of stochastic disturbances from profit maximization.

[7] Since all derived parameters are linear combinations of the least-squares estimators, the variances of the derived parameters are obtained straightforwardly.

Assuming that the elements of the regressor matrix are predetermined variables, application of IZEF to (14) yields consistent and asymptotically efficient estimates of α_L, α_K, γ_{LL}, γ_{LK}, and γ_{KK}. All of the remaining α and γ estimates, which are linear combinations of these estimates, also have these desirable asymptotic properties. The assumption that the input ratios are predetermined in (14) is probably not strictly tenable. For if input prices are exogenous, the observed input ratios are behaviorally determined and endogenous. An iterative three-stage least-squares estimation procedure might therefore be more appropriate, but we have not employed it.[8]

B. Measurement of Variables

Assume that the production function (1) may be written as $q = f(K, L, N)$ and that the distributive shares of these inputs exhaust total cost. The total annual costs of production are thus apportioned among the wage bill, gross quasi-rents, and the cost of natural resource products. Total cost is computed, in each industry, by total labor costs (the wage bill) plus total capital costs (the sum of these two components being equal to value added) plus the direct costs of natural resource products. The distributive shares are calculated by dividing the cost attributable to each input by total cost.

For purposes of estimation, the 235 four-digit manufacturing industries that incur any direct cost of natural resource products are classified according to the two-digit product group to which they belong. It is assumed that the production function parameters are constant and equal among subindustries within a specific two-digit product group, but they may differ across two-digit categories. With this scheme of classification, sufficient observations are obtained for estimation in 12 distinct product groups. The last is a residual category of all four-digit industries which could not be used for estimation within a specific two-digit product group.

---

[8] Berndt and Christensen (1973*b*) consider in an aggregative time-series context the prospect that inputs are determined simultaneously with output and relative shares. They accordingly employ both IZEF and iterative three-stage least squares to estimate translog parameters for U.S. manufacturing as an aggregate. Their IZEF and iterative three-stage least-squares estimates are practically identical.

TABLE 1

COST SHARES OF RESOURCE PRODUCTS IN TWO-DIGIT PRODUCT GROUPS, 1963

Product Group	Average Resource Cost Share
Food and beverages (20)	.371
Textiles and apparel (22 and 23)	.118
Lumber products (24)	.258
Pulp and paper products (26)	.102
Chemical products (28)	.039
Stone, clay, and glass (32)	.109
Primary metals (33)	.152
Fabricated metals (34)	.001
Nonelectrical machinery (35)	.001
Electrical machinery (36)	.001
Transportation equipment (37)	.001
All others (19, 25, 27, 30, 31, 38, and 39)	.009

SOURCE.—See Appendixes A and B.

There is considerable variation among individual industries and across product groups in the average share of total cost attributable to natural resource products. The average relative share of resource product cost in each two-digit category is given in table 1. In product groups 34–37, natural resource products account on the average for only one-tenth of 1 percent of total costs, whereas in product groups 20, 22–24, 26, 32, and 33 they are significant components of industry costs.

With regard to (14), the labor–resource product ratio is calculated, in each four-digit industry, as man-years of employment divided by a quantity index of resource products used.[9] The capital–resource product ratio is computed, in each industry, by dividing the gross book value of the industry capital stock by this index of natural resource products. The book value of an industry's capital stock is quite distinct from a flow of homogeneous capital services; and the use of reported capital stocks to estimate capital services across industries is a practice subject to serious limitations, of which we mention only three. (i) The physically heterogeneous capital stocks were undoubtedly accumulated within the various industries throughout somewhat different time periods and therefore at different average price levels. Only if the average rate of capital-embodied technological progress were equal to the average rate of inflation in capital goods prices, within in each industry, would the book values reflect effective capital stocks comparable across industries. (ii) There are probably unknown interindustry differences in the percentage of reported gross book values remaining economically useful. (iii) There are also probably interindustry differences in capacity utilization rates of the reported book values. One should imagine (ii) and (iii) to be present more

---

[9] The construction of this index, which was prepared by Humphrey, is discussed in detail in Appendix A.

TABLE 2

ESTIMATES OF TRANSLOG PARAMETERS IN TWO-DIGIT PRODUCT GROUPS, 1963

Product Group	$\hat{\alpha}_L$	$\hat{\alpha}_K$	$\hat{\gamma}_{LL}$	$\hat{\gamma}_{LK}$	$\hat{\gamma}_{KK}$	N
20	−.102	.417*	.055*	−.027	.066*	64
	(1.13)	(4.19)	(2.84)	(1.42)	(3.15)	
22 and 23	−.196	.272	.068*	−.025	.066*	36
	(1.48)	(1.70)	(3.56)	(1.23)	(2.64)	
24	−.137	.775*	.112*	−.085*	.117*	20
	(0.95)	(5.41)	(4.11)	(3.14)	(4.19)	
26	.881*	−.368	−.232	.315*	−.384*	14
	(2.27)	(0.89)	(2.04)	(2.42)	(2.53)	
28	.588*	.238*	−.043	.023	.029	24
	(4.47)	(2.19)	(1.69)	(0.87)	(0.93)	
32	−.414*	.798*	.156*	−.116*	.152*	48
	(5.47)	(10.14)	(10.44)	(6.63)	(6.21)	
33	−.303*	.627*	.135*	−.103*	.139*	36
	(2.91)	(6.16)	(5.63)	(3.97)	(4.82)	
34	.076	.913*	.091	−.091	.092	34
	(0.25)	(3.04)	(1.50)	(1.49)	(1.51)	
35	.724*	.263	−.039	.040	−.040	56
	(2.83)	(1.02)	(0.74)	(0.76)	(0.76)	
36	.125	.858*	.071	−.071	.073	46
	(0.52)	(3.54)	(1.51)	(1.50)	(1.54)	
37	−.302	1.291*	.174*	−.174*	.174*	20
	(1.03)	(4.40)	(3.09)	(3.08)	(3.09)	
19, 25, 27, 30, 31, 38, and 39	.156	.721*	.062*	−.057*	.063*	72
	(1.11)	(5.36)	(2.53)	(2.26)	(2.29)	

NOTE.—t-statistics, in parentheses below the estimated coefficients, are t-ratios for the hypothesis that the parameter differs from zero.

* Coefficient is significantly different from zero at $P \leq .05$.

frequently among relatively older industries. There seem to be no strongly defensible rules of thumb for adjusting the book values; we proceed in the knowledge of their shortcomings.

IV. Estimation

A. Translog Profit-maximizing Equations

Parameters estimated for equation (14) appear in table 2. There are no a priori restrictions concerning the signs of γ_{LL}, γ_{LK}, and γ_{KK}, so two-tail tests of significance are appropriate. We select $P \leq .05$ as the critical level of significance.

We first test for stability of the input demand equations. In five groups, 34–37 and "all others" (last row), the parameter estimates are such that the bordered principal minors of $|F|$ do not satisfy the required inequalities, thereby implying that the underlying input demand equations are unstable. The principal minors are evaluated by inserting the parameter estimates and the data sample means in (11a). Thus, the principal minors are themselves random variables; but they are highly complex, nonlinear combinations of the parameter estimates, and it is not feasible to obtain even approximate standard errors for them.

Although several parameter estimates in these five groups are significant in the statistical sense, the entire set in each group must be rejected as implausible from the standpoint of economic theory. Common sense dictates that the data must be generated by stable input demand equations associated with some unknown production function, so the observed data are simply incompatible with the joint hypotheses of profit maximization subject to a translog production specification. We attribute this finding to the fact that, because such a small amount of natural resource products is used in these groups (in each, less than 1 percent in terms of total costs), the translog function cannot reliably be applied to explain the input relations. Reference to equation (14) indicates more precisely the nature of the statistical problem. If natural resource products account for a meager fraction of total costs, then for the regression in question, the sum of M_L and M_K essentially exhausts the distributive cost shares. And the variation in these shares is not systematically related, in an economic sense, to the variation in the (L/N) and (K/N) ratios, since resource products are practically irrelevant as inputs.

Consider now the product groups with stable input demand functions. A second hypothesis of interest is that the underlying production function has the simple Cobb-Douglas form. If we maintain the symmetry and linear homogeneity assumptions, the appropriate statistical test is $\gamma_{LL} = \gamma_{LK} = \gamma_{KK} = 0$. We thus apply the IZEF method to a version of equation (14) in which the input ratios are excluded as regressors. We then construct F-ratios to determine whether their exclusion significantly increases the residual variance. In all groups except 26 there is a significant increase in residual variance at $P \leq .05$, so the Cobb-Douglas form is rejected.[10]

Using the estimated parameters in table 2, together with estimates of the remaining parameters implied by the restrictions $\gamma_{ij} = \gamma_{ji}$ $(i \neq j)$ and constant returns to scale, we compute for each product group the elements of $|F|$ and evaluate $|F_{KL}|$, $|F_{LN}|$, and $|F_{KN}|$. Using (11) we then calculate the Allen partial elasticities of substitution, σ_{KL}, σ_{LN}, and σ_{KN}, which are given in table 3. Recall that in general these partial elasticities are variable, so these estimates are evaluated at sample means.[11]

---

[10] See Solow (1955–56) for discussion of functional separability in the context of the Cobb-Douglas function. We note here that, according to an alternative test, utilizing a translog cost function, the Cobb-Douglas function is rejected in product group 26 (see table 6).

[11] Using equation (5), the mean value of the relative share of labor within a product group is $M_L^* = \alpha_L + \sum_j \gamma_{Lj} \ln X_j^*$ (for $j = K, L, N$). Thus, as an illustrative calculation, the first- and second-order derivatives with respect to labor in equations (2)–(4), evaluated at sample means, are

$$f_L = (q^*/X_L^*)M_L^*, \tag{2'}$$

$$f_{LL} = (q^*/X_L^{*2})(\gamma_{LL} + M_L^{*2} - M_L^*), \tag{3'}$$

$$f_{LK} = (q^*/X_L^* X_K^*)(\gamma_{KL} + M_K^* M_L^*), \tag{4'}$$

where starred values refer to sample means.

TABLE 3

PARTIAL ELASTICITIES OF SUBSTITUTION IN TWO-DIGIT PRODUCT GROUPS, 1963

PRODUCT GROUP	TRANSLOG PRODUCTION FUNCTION σ_{KL}	σ_{LN}	σ_{KN}
20	1.44	1.51	1.34
22 and 23	0.76	4.95	5.03
24	2.54	1.24	1.73
26	0.37	2.46	0.50
28	36.75	−446.35	666.71
32	2.29	4.89	4.24
33	2.35	1.88	2.40

All point estimates of σ_{KL} are positive, as one might expect theoretically. Notice that the point estimates of σ_{LN} and σ_{KN} are strikingly similar in five of the seven product groups. It is also noteworthy that within most of these resource-intensive sectors the substitutability between capital and resource products as well as that between labor and resource products appears to be at least as high as the substitutability between capital and labor.

Consider now the problem of conducting statistical tests of significance. Equations (11) and (11a), which are themselves expressions in terms of the parameters appearing in the first and second derivatives of (1), show that each partial elasticity is a *nonlinear function* of *all* of the estimated α and γ coefficients. Indeed, the explicit solution for each partial elasticity in terms of the estimated regression coefficients is an exceedingly complex, nonlinear expression involving 70 terms.[12] Hence, it is not feasible to obtain approximate standard errors of the individual partial elasticities from a knowledge of the estimated variances and covariances of the regression coefficients. For this reason we cannot test with these estimates the hypotheses of strong complementarity, that is, the hypotheses $\sigma_{LN} \leq 0$, $\sigma_{KN} \leq 0$.[13]

The hypothesis $\sigma_{KN} < \sigma_{LN}$ ($\sigma_{LN} < \sigma_{KN}$) is termed *relative* complementarity of capital (labor) and resource products compared with labor (capital) and resource products. As shown in Section II, $\sigma_{KN} = \sigma_{LN}$ is equivalent to functional separability of K and L from N. Thus, one may test $\sigma_{KN} = \sigma_{LN}$ against the two-tail alternative by testing the hypothesis $M_K\gamma_{LN} - M_L\gamma_{KN} = 0$. We construct in each product group a test statistic $M_K\hat{\gamma}_{LN} - M_L\hat{\gamma}_{KN}$, using the sample means of the relative shares

---

[12] This expression is a mathematically straightforward, but tedious, expansion of equation (11), and there is no need to reproduce it here. It is available from the authors on request.

[13] Estimates of the standard errors of the partial elasticities associated with a translog cost function are obtained and the hypotheses of strong complementarity are tested in Section V.

TABLE 4

TESTS OF RELATIVE COMPLEMENTARITY: $\sigma_{KN} = \sigma_{LN}$

Product Group	$\bar{M}_K\hat{\gamma}_{LN} - \bar{M}_L\hat{\gamma}_{KN}$	t
20	.0028 (.0026)	1.06
22 and 23	−.0003 (.0071)	0.04
24	−.0052 (.0029)	1.79
26	.0691 (.0226)	3.06
28	−.0285 (.0095)	3.02
32	.0012 (.0081)	0.15
33	−.0029 (.0043)	0.68

NOTE.—Approximate standard errors of $(\bar{M}_K\hat{\gamma}_{LN} - \bar{M}_L\hat{\gamma}_{KN})$ are given in parentheses.

of capital and labor, together with the linear restrictions

$$\hat{\gamma}_{LN} = (-\hat{\gamma}_{LK} - \hat{\gamma}_{LL}) \quad \text{and} \quad \hat{\gamma}_{KN} = (-\hat{\gamma}_{KK} - \hat{\gamma}_{KL}),$$

associated with equation (1).[14]

The test results appear in table 4. The hypothesis $\sigma_{KN} = \sigma_{LN}$ is rejected in two of the seven product groups. In the pulp and paper product group (SIC 26), natural resource products appear to be more highly substitutable with labor than with capital, while the converse holds among the chemical industries (SIC 28). The major conclusion to be drawn from these tests is that within most of the resource-intensive groups capital and labor are equally facile substitutes for natural resource products. The viewpoint of predominant (relative) capital–resource product complementarity receives no support.

B. Input Relations in Resource-intensive and Resource-sparse Sectors

It was noted above that the factor demand stability conditions are not met in any of the five product groups in which natural resource products

---

[14] Since the derived estimates $\hat{\gamma}_{LN}$ and $\hat{\gamma}_{KN}$ are linear combinations of the regression coefficients whose asymptotic variances and covariances are known,

$$\begin{aligned} &\text{var}\,(M_K\hat{\gamma}_{LN} - M_L\hat{\gamma}_{KN}) \\ &\doteq (M_K)^2\,\text{var}\,(\hat{\gamma}_{LL}) + (M_L - M_K)^2\,\text{var}\,(\hat{\gamma}_{KL}) + (M_L)^2\,\text{var}\,(\hat{\gamma}_{KK}) \\ &\quad - 2M_K(M_L - M_K)\,\text{cov}\,(\hat{\gamma}_{LL}, \hat{\gamma}_{KL}) \\ &\quad - 2M_KM_L\,\text{cov}\,(\hat{\gamma}_{LL}, \hat{\gamma}_{KK}) \\ &\quad + 2(M_L - M_K)M_L\,\text{cov}\,(\hat{\gamma}_{KL}, \hat{\gamma}_{KK}), \end{aligned}$$

where var ( ) and cov ( ) refer to estimated asymptotic values. Using arithmetic means of the observed distributive shares M_L and M_K, the test statistic $(\bar{M}_K\hat{\gamma}_{LN} - \bar{M}_L\hat{\gamma}_{KN})/(\text{var}\,[\bar{M}_K\hat{\gamma}_{LN} - \bar{M}_L\hat{\gamma}_{KN}])^{1/2}$ will have approximately the t distribution.

account for less than 1 percent of total costs. And the elasticity estimates for the chemical products group, in which natural resource products constitute less than 5 percent of costs, seem implausibly large. To establish a sharper distinction between these groups and those that use resource products more intensively, we divided the 12 groups into two sectors, according to whether the average resource cost share exceeded or fell short of 5 percent. Equation (14) was then estimated by IZEF using these alternative samples, as well as an aggregate sample covering all industries. The estimates appear in table 5. One may test for equality of the two sectoral regression equations using an *F*-test suggested by Chow (1960).[15] The computed *F* is 4.316, a value leading to rejection of the hypothesis that the parameters are equal (see the note in table 5). Furthermore, the parameter estimates for the resource-intensive sector I satisfy the input demand stability conditions, but those for the resource-sparse sector II do not. We regard these findings, together with the individual product group estimates, as strong confirmation that the translog production model cannot fruitfully be applied in those sectors that make only minor use of natural resource products. But for the broader purposes of our study, and for social policy, these sectors are far less important than those making relatively intense use of resource products. Accordingly, in the remainder of the paper, attention is centered on the resource-intensive product groups.

The partial elasticities of substitution within the two sectors, evaluated at sample means, also appear in table 5. The estimates for sector I indicate that the pairwise substitutability among capital, labor, and natural resource products is practically identical. We conduct tests of the null hypotheses $\sigma_{KL} = \sigma_{LN}$, $\sigma_{KL} = \sigma_{KN}$, and $\sigma_{KN} = \sigma_{LN}$. In each instance the hypothesis of equality cannot be rejected at $P \leq .05$. One must, of course, not ascribe any economic significance to the estimated elasticities in sector II, since the factor demand stability conditions are not satisfied.

V. An Alternative Approach to Input Substitution

One may alternatively investigate substitution technologies by means of cost functions, as has been shown by Shephard (1953, 1970), Uzawa (1962), and Diewert (1971).

Assume that economic units produce alternative rates of output according to a cost function

$$C = C(Q, P_i), \qquad i = 1, \ldots, n, \tag{15}$$

---

[15] This test relies on the maintained hypothesis that the disturbance vectors associated with the two sectoral regressions are asymptotically independent.

TABLE 5

Translog Production Parameters for Resource-intensive and Resource-sparse Sectors

	$\hat{\alpha}_L$	$\hat{\alpha}_K$	$\hat{\gamma}_{LL}$	$\hat{\gamma}_{LK}$	$\hat{\gamma}_{KK}$	SSE	N	σ_{KL}	σ_{LN}	σ_{KN}	Average Resource Cost Share
Sector I (20, 22–24, 26, 32, 33)	−.123* (2.19)	.579* (10.56)	.085* (7.63)	−.058* (5.11)	.082* (6.75)	3.29160	218	1.59	1.53	1.38	.208
Sector II (28, 34–37, "all others")	−.054 (0.73)	.910* (12.43)	.095* (6.60)	−.084* (5.57)	.084* (4.99)	2.17071	252	1.49	−5.63	2.38	.007
Sectors I and II	−.138* (3.25)	.631* (14.52)	.089* (10.30)	−.065* (6.94)	.086* (8.27)	5.71857	470	1.45	2.18	1.91	.100

NOTE.—SSE denotes the sum of squares of residuals in the respective regressions. An *F*-test for equality between regression equations in Sectors I and II is based on the test statistic

$$\hat{F} = \frac{[SSE_{(I+II)} - SSE_{(I)} - SSE_{(II)}]/5}{[SSE_{(I)} + SSE_{(II)}]/460} = 4.316.$$

The critical F for rejecting equality at $P \leq .01$ is approximately 3.08.

where C and Q are total costs and output and P_i is the price of the ith input. Assume further that C is (i) a positive, real-valued continuous function in Q that tends to infinity as Q tends to infinity; (ii) linearly homogeneous in P_i; and (iii) a concave function in P_i for all positive rates of output.[16] It has been shown by Shephard (1953, 1970) and Diewert (1971) that (iv) if C is continuously differentiable with respect to P_i, and is minimized for all $Q > 0$ by producers who purchase specifiable inputs on competitive markets, there exists a well-defined production function that is a duality relation to the cost function (15). The production function may not be expressible in explicit, parametric form; but the technologies of input substitution (as well as the returns to scale and other properties) may be determined by analysis of the cost function alone.

Let us assume that there exists a minimum cost function for alternative rates of output within each of the resource-intensive product groups. As before, the specifiable factors of production are labor (L), capital (K), and natural resource products (N). One must now specify a parametric form for the cost function, more specifically a form that permits estimation of the partial substitution elasticities. A translog form, suggested originally by Christensen et al. (1970), seems highly suitable for two reasons: it permits direct estimation of substitution elasticities, and it entails no a priori restrictions respecting either their values or constancy. It should perhaps be emphasized that the production function underlying the translog cost function is not necessarily of the translog type analyzed in Section II. One knows only that, if assumptions (i)–(iv) are applicable within the "economic region" of the translog cost function, there is some production function, whose form is not explicit, associated with the cost function. One can nonetheless estimate its substitution elasticities from certain parameters of the cost function.

A translog cost function may be expressed as

$$\begin{aligned}\ln C = a_0 + \theta_1 \ln Q + \tfrac{1}{2}\theta_2 (\ln Q)^2 + \sum_i \beta_i \ln P_i + \tfrac{1}{2} \sum_i \sum_j \varepsilon_{ij} \\ \times \ln P_i \ln P_j + \sum_i \delta_i \ln P_i \ln Q, \qquad i = K, L, N, \qquad (16)\end{aligned}$$

$$\varepsilon_{ij} = \varepsilon_{ji} \qquad \text{for } i \neq j.$$

The property of linear homogeneity in factor prices imposes the following restrictions on the parameters of (16):

$$\sum_i \beta_i = 1, \qquad (17a)$$

$$\sum_j \varepsilon_{ij} = \sum_i \varepsilon_{ji} = \sum_i \sum_j \varepsilon_{ij} = 0, \qquad i,j = K, L, N, \qquad (17b)$$

$$\sum_i \delta_i = 0. \qquad (17c)$$

---

[16] Assumption (i) incorporates Shephard's propositions Q. 2, Q. 7, and Q. 9; assumption (ii) corresponds to Shephard's Q. 3; and assumption (iii) corresponds to Shephard's Q. 6. See Shephard (1970, p. 83); for comparison see Diewert (1971, pp. 489–90).

The property that C is a continuously increasing function of factor prices is expressed by three equations:

$$\frac{\partial \ln C}{\partial \ln P_i} = \beta_i + \sum_j \varepsilon_{ij} \ln P_j + \delta_i \ln Q, \qquad i, j = K, L, N. \tag{18}$$

To develop equations (18) as stochastic estimating forms, one may invoke Shephard's lemma (1970, p. 171): *Along the minimum cost expansion path the equilibrium employment of the ith input is*

$$X_i^*(Q, P_i) = \partial C/\partial P_i, \qquad i = K, L, N. \tag{19}$$

The relative shares in total cost are, by definition,

$$M_i = P_i X_i / C. \tag{20}$$

Substituting X_i^* from (19) into (20), relative shares on the cost-minimizing expansion locus are

$$M_i^* = (\partial C/\partial P_i)(P_i/C) \equiv \partial \ln C/\partial \ln P_i, \qquad i = K, L, N. \tag{21}$$

Accordingly, M_i^* may be substituted for $(\partial \ln C/\partial \ln P_i)$ in equations (18). There will, of course, be some discrepancies between the minimum-cost shares, M_i^*, and the observed shares, M_i, attributable to entrepreneurial errors, inertia, and other constraints preventing the exact minimization of costs. Thus, $M_i = M_i^* + e_i$, where e_i is the disturbance from cost minimization. Substituting these expressions for observed shares in equations (18) yields three stochastic cost-minimizing equations:

$$M_i = \beta_i + \sum_j \varepsilon_{ij} \ln P_j + \delta_i \ln Q + e_i, \qquad i, j = K, L, N. \tag{22}$$

The restriction of linear homogeneity in factor prices, equations (17a)–(17c), shows that parameter estimates of any two equations exactly identify all parameters of the three-equation system (22). Furthermore, unrestricted estimation of each of the three equations (22) does not ensure that $\varepsilon_{ij} = \varepsilon_{ji}$ $(i \neq j)$. One is thus confronted with the same type of identification problem as encountered in the production function context. And a similar method of restricted estimation is used to solve it: substitute the restrictions $\varepsilon_{LN} = \varepsilon_{NL} = (-\varepsilon_{LL} - \varepsilon_{KL})$ and $\varepsilon_{KN} = \varepsilon_{NK} = (-\varepsilon_{KK} - \varepsilon_{KL})$ in the first two of equations (22) to obtain the regression model

$$\begin{bmatrix} M_L \\ M_K \end{bmatrix} = \begin{bmatrix} 1 & 0 & \ln\left(\frac{P_L}{P_N}\right) & \ln\left(\frac{P_K}{P_N}\right) & 0 & \ln Q & 0 \\ 0 & 1 & 0 & \ln\left(\frac{P_L}{P_N}\right) & \ln\left(\frac{P_K}{P_N}\right) & 0 & \ln Q \end{bmatrix} \begin{bmatrix} \beta_L \\ \beta_K \\ \varepsilon_{LL} \\ \varepsilon_{KL} \\ \varepsilon_{KK} \\ \delta_L \\ \delta_K \end{bmatrix} + \begin{bmatrix} e_L \\ e_K \end{bmatrix}. \tag{23}$$

TABLE 6

PARTIAL ELASTICITIES OF SUBSTITUTION ESTIMATED FROM TRANSLOG COST FUNCTION, 1963

Product Group	$\bar{\sigma}_{KL}$	$\bar{\sigma}_{LN}$	$\bar{\sigma}_{KN}$	N
20	0.698*† (.282)	0.641* (.201)	0.641*† (.154)	64
22 and 23	1.099* (.199)	0.473 (.298)	0.167† (.261)	36
24	2.548*† (.678)	0.606 (.328)	1.119* (.397)	20
26	0.534*† (.073)	4.492*† (.505)	−0.684† (.437)	14
32	1.659*† (.296)	0.064 (1.083)	−0.580 (.823)	48
33	0.726* (.353)	5.918*† (1.138)	1.841 (.994)	36

SOURCE.—See Appendixes A and B.
NOTE.—Estimated asymptotic standard errors are in parentheses.
* Denotes elasticities significantly different from zero at $P \leq .05$.
† Denotes elasticities significantly different from 1 at $P \leq .05$.

Uzawa (1962) showed that the partial elasticity of substitution between inputs i and j is expressible in terms of derivatives of the cost function as

$$\sigma_{ij} = C(\partial^2 C/\partial P_i\, \partial P_j)/(\partial C/\partial P_j)(\partial C/\partial P_i). \tag{24}$$

Direct substitution of the parameters from (23) in (24) yields

$$\sigma_{ij} = (\varepsilon_{ij}/M_i M_j) + 1, \qquad i \neq j. \tag{25}$$

These partial elasticities generally vary with the values of relative input shares (unless the product $M_i M_j$ in [25] is constant). A set of average values may nonetheless be computed at the sample means:

$$\bar{\sigma}_{ij} = (\hat{\varepsilon}_{ij}/\bar{M}_i \bar{M}_j) + 1. \tag{26}$$

Their variances are estimated from

$$\text{var}\,(\bar{\sigma}_{ij}) = (1/\bar{M}_i \bar{M}_j)^2 \,\text{var}\,(\hat{\varepsilon}_{ij}). \tag{27}$$

The parameters of relation (23) are estimated by the IZEF method for each of the resource-intensive sectors. The estimated mean elasticities, together with their estimated standard errors, appear in table 6. For each product group, the comprehensive fixed-coefficients hypothesis $\sigma_{KL} = \sigma_{LN} = \sigma_{KN} = 0$ is rejected because at least one partial elasticity is significantly different from zero. The Cobb-Douglas hypothesis $\sigma_{KL} = \sigma_{LN} = \sigma_{KN} = 1$ is also rejected in all groups.

Consider now the individual estimates. In each group, σ_{KL} is significantly positive, and in two groups it is significantly greater than 1. Three of the σ_{LN} estimates are statistically significant, but in three groups the strong complementarity hypothesis $\sigma_{LN} \leq 0$ cannot be rejected. Finally, only two of the σ_{KN} estimates are significantly positive, although the estimate in group 33 is near the borderline of significance. The hypothesis of strong capital–natural resource product complementarity is thus sustained in four of these resource-intensive sectors.

As in the production function context, we test for equality of the substitution elasticities σ_{KN} and σ_{LN}. In two of the six product groups, SIC 26 and 33, σ_{KN} is significantly smaller than σ_{LN}; thus, there is some evidence of relative capital-resource complementarity in the cost function models, as was found in SIC 26 with the production models.

One might naturally want to compare the substitution elasticities obtained from a translog cost function with those estimated from a translog production function. Some words of caution are in order before making such comparisons, because the affinities between the two modes of investigation are perhaps more apparent than real. Indeed, the two methods may appear to be deceptively similar, since the same observed cost shares are used empirically as dependent variables in the alternative estimating equations (14) and (23). Nonetheless, the two methods are basically alternative frameworks for research.

First, and most important, the theoretical assumptions underlying the two approaches to input substitution are distinctly different. The production model incorporates as a maintained hypothesis that technology is approximately describable by a linearly homogeneous translog production function form. By contrast, the translog cost model is based on the assumption that there exists *some* production function whose explicit form is unspecified (and need not for analytical purposes be specified) but whose substitution elasticities can be ascertained from certain parameters of a (minimum) cost function. On theoretical grounds, then, one might expect the similarity of substitution elasticities obtained from the two models to depend both on the closeness with which the translog functional form approximates the true, but unknown, production function and on the magnitude of disturbances in the minimization of costs.

There is, second, a potential errors-in-variables problem. Recall that the natural resource product quantity indices are computed by dividing the value flows of natural resource products by the natural resource product price indices. So upward or downward biases in the resource product price indices impart alternately downward or upward biases in the resource product quantity indexes. Hence, even if the translog form were a very close approximation, measurement errors in regressors would cause some discrepancies in the estimated elasticities. (The closer similarity between the alternative estimates of σ_{KL} indeed suggests that measurement errors primarily affect the σ_{LN} and σ_{KN} estimates.)

Finally, one may recall that the average elasticities calculated for the translog production model are complex, nonlinear functions of all the production parameter estimates (see eqq. [11] and [11a]). By contrast, the elasticities estimated from the translog cost function are simple, linear combinations of the estimated price response coefficients of the cost function. It is well known that a parameter estimate obtained as a nonlinear combination of consistent estimators is itself a consistent estimator. But it is not in general unbiased. Hence, even if the two sources of discrepancy mentioned above were unimportant, there would be some differences between the expected values of the alternative estimates of the substitution elasticities.

With these caveats in mind, one may proceed to some gingerly comparisons. Aligning the estimates in table 6 with the corresponding estimates in table 3 reveals that the capital-labor elasticities are quite similar in four sectors. By contrast, the σ_{LN} cost function estimate is markedly lower in four sectors than the corresponding production function estimate. Finally, in each of the six product groups, the σ_{KN} cost function estimate is lower than the corresponding production function estimate. The cost function estimates suggest somewhat higher labor–resource product substitutability than capital–resource product substitutability. And these results indicate that the natural-resource-based constraints to growth among these sectors may be difficult to offset simply through the more intensive use of capital.

VI. Summary and Conclusions

This paper investigates factor substitution using two alternative research designs. Both the production and the cost function frameworks are static. Accordingly, the interindustry differences in factor shares within product groups are attributed exclusively to input substitution among the broad classes of reproducible capital, labor, and natural resource products. To the extent that such differences in relative shares are ascribable in part to nonneutral differences in production technologies, our estimated substitution elasticities may either understate or overstate the true substitutability along static isoquant surfaces.[17] But whether our estimates do or do not incorporate some effects of nonneutral technology is in a way unimportant from the broader viewpoint of evaluating the resource-constrained limits to long-run growth. For regardless of the underlying causes of input substitution (that is, whether it is technology induced or price induced), the evidence suggests that labor and, to a lesser degree, capital are substitutable for natural resource products among most of the resource-using product groups of American manufacturing.

With the exception of the pulp and paper sector (SIC 26), there is little evidence that labor and capital differ greatly in their facility of substitution for resource products. Estimates from the cost model indicate that labor is the better substitute within product groups 26 and 33. Within the other product groups, both the production and the cost function estimates imply that capital and labor may be analyzed as functionally separable from natural resource products. These results provide only limited support for the view that relative capital and natural resource complementarity accounts for Leontief's paradoxical findings concerning the commodity composition of American foreign trade. It would seem that more fruitful explanations for the Leontief paradox may be found in the patterns of tariff protection (Travis 1964) and international differences in the commodity content of human capital (Keesing 1965; Kenen 1965).

---

[17] The same problem is encountered if one estimates the simple elasticity of substitution between capital and labor according to the assumption of Hicks-neutral technological change. For discussion of this problem from the standpoint of statistical specification error, see Moroney (1972, p. 73–82, 153–54).

Finally, the eleventh-hour predictions of secular stagnation, based on the joint propositions of impending resource scarcity and fixed coefficients between natural resource products and other inputs, appear too pessimistic. Even our minimum-substitution estimates imply that future demands for natural resources should be mitigated by the increased use of labor or capital within manufacturing. Additional succor within the domestic economy may be found in two areas: (i) by direct substitution of capital and labor for basic natural resources in the production of natural resource products and (ii) by resource-saving changes in the composition of final demand.

The detailed and important questions concerning substitutability among labor, capital, and highly specific types of resource products remain to be answered. The study of structural change in input usage within industries over time would seem to be a particularly fruitful area for future research. The frameworks used in this paper could readily be modified to account for factor-biased technological change in such applications.

Appendix A

Originally, data were collected for 291 four-digit SIC U.S. manufacturing industries for 1963. The number of industries was reduced to 235 by deleting those industries that did not use any natural resource products as inputs. As discussed in the text, the translog production function cannot be applied to industries that use nonpositive quantities of any specifiable input. Price (wage rates), quantity (man-years of employment), and distributive shares (the wage bill divided by value added plus the value of resource products) for labor inputs are taken or computed from U.S. Bureau of the Census (1966, table 3, pp. 46–67). It is assumed that returns to labor and capital exhaust value added. Consequently, the cost of capital is computed as gross quasi-rent (value added minus the wage bill) divided by the gross book value of the capital stock (from U.S. Bureau of the Census [1968, chap. 7, pp. 150–71]). Gross book value of the capital stock is used as a proxy for an unavailable measure of the "quantity" of capital.

Input value data for the 19 renewable and nonrenewable natural resource product inputs into the 235 manufacturing industries that utilize some natural resource products are taken from the detailed (360 × 360) value flows of the

Office of Business Economics (1969, input-output table). Renewable natural resource products comprise sector numbers 1.01–1.03, 2.01–2.07, 3.00, and 20.01. Nonrenewable resource products are sectors 5.00, 6.01–6.02, and 7.00–10.00. Sector 4.00, services, was deleted. Detailed price data for the seven nonrenewable natural resource products were kindly made available to us by Mr. Louis J. Owen of the Industry Division, U.S. Bureau of the Census. These prices were originally used to generate the published sector unit value indices in U.S. Bureau of the Census and Board of Governors of the Federal Reserve System (1970, table 5, pp. 8–9). The two- and three-digit price data we used represent value-weighted averages of the above detailed five- and seven-digit price data. In this manner we obtained prices for the seven nonrenwable natural resource products. These prices were then weighted by nonrenewable input-output value flows to the 235 natural-resource-using industries, giving 235 average prices of nonrenewable resource inputs.

Prices for the remaining 12 renewable natural resource products represent our own computations of the major output components of these sectors as published in the Census of Agriculture, U.S. Bureau of the Census (1966, chap. 6, pp. 601–2), and selected additional sources for particular output groups (e.g., vegetables, feed grains, and logging camps, among others). These prices refer to 1964, but we used them with our other statistics pertaining to 1963. Mr. C. A. Stokstad of the Statistical Reporting Service, U.S. Department of Agriculture, was very helpful in providing certain commodity classifications and information concerning average commodity prices. Once the 12 renewable resource product prices were obtained, they were weighted by renewable input-output value flows into the 235 natural-resource-using sectors, thereby yielding average prices of renewable resource inputs. It was then a straightforward matter to compute the 235 value-weighted average prices of all renewable and nonrenewable resource products used by the 235 manufacturing industries. Quantities of all natural resource inputs into these industries were estimated by deflating the value of resource products by our computed average price.

Appendix B

Nineteen Natural Resource Products Used in This Study

Twelve renewable natural resource products (sector number in parentheses):

Dairy farm products (1.01)
Poultry and eggs (1.02)
Meat, animals, and miscellaneous products (1.03)
Cotton (2.01)
Food feed grains and grass seeds (2.02)
Tobacco (2.03)
Fruits and tree nuts (2.04)
Vegetables, sugar, and miscellaneous crops (2.05)
Oil-bearing crops (2.06)
Forest, greenhouse, and nursery products (2.07)
Forestry and fishery products (3.00)
Logging camps and logging contractors (20.01)

Seven nonrenewable natural resource products:
Iron and ferroalloy ores mining (5.00)
Copper ore mining (6.01)
Nonferrous metal ores mining, except copper (6.02)
Coal mining (7.00)
Crude petroleum and natural gas (8.00)

Stone and clay mining and quarrying (9.00)
Chemical and fertilizer mineral mining (10.00)

NOTE.—Sector numbers are from Office of Business Economics (1969).

References

Adler, John H. "Changes in the Role of Resources at Different Stages of Economic Development." In *Natural Resources and Economic Growth,* edited by J. J. Spengler. Washington: Resources for the Future, 1961.

Allen, R. G. D. *Mathematical Analysis for Economists.* London: Macmillan, 1938.

Baldwin, R. E. "Determinants of the Commodity Structure of U.S. Trade." *A.E.R.* 61 (March 1971): 126–46.

Barnett, Harold, and Morse, Chandler. *Scarcity and Growth.* Baltimore: Johns Hopkins Press for Resources for the Future, 1963.

Behrens, William W. "The Dynamics of Resource Utilization." Unpublished. System Dynamics Group, Massachusetts Inst. Tech., May 1971.

Berndt, E. R., and Christensen, L. R. "The Internal Structure of Functional Relationships: Separability, Substitution, and Aggregation." *Rev. Econ. Studies* 40 (July 1973): 403–10. (*a*)

———. "The Translog Function and the Substitution of Equipment, Structures, and Labor in U.S. Manufacturing, 1929–1968." *J. Econometrics* 1 (March 1973): 81–114. (*b*)

Carter, Anne P. *Structural Change in the American Economy.* Cambridge, Mass.: Harvard Univ. Press, 1970.

Chow, G. C. "Tests of Equality between Sets of Coefficients in Two Linear Regressions." *Econometrica* 28 (July 1960): 591–605.

Christensen, L. R.; Jorgenson, D. W.; and Lau, L. "Conjugate Duality and the Transcendental Logarithmic Production Function." Unpublished paper presented at the Second World Congress of the Econometric Society, Cambridge, 1970.

Diewert W. E. "An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function." *J.P.E.* 79, no. 3 (May/June 1971): 481–507.

Georgescu-Roegen, Nicholas. *The Entropy Law and the Economic Process.* Cambridge: Mass.: Harvard Univ. Press, 1971.

Hagen, E. "Social Accounts and the Incremental Capital-Output Ratio in Underdeveloped Countries." Paper presented at the Third Conference of the International Association for Research in Income and Wealth, Castelgandolfo, 1953.

Keesing, D. B. "Labor Skills and International Trade: Evaluating Many Trade Flows with a Single Measuring Device." *Rev. Econ. and Statis.* 47 (August 1965): 287–94.

Kenen, P. B. "Nature, Capital, and Trade." *J.P.E.* 73, no. 5 (October 1965): 437–60.

Kmenta, J., and Gilbert, R. "Small Sample Properties of Alternative Estimators of Seemingly Unrelated Regressions." *J. American Statis. Assoc.* 63 (December 1968): 1180–2000.

Leontief, W. W. "Introduction to a Theory of the Internal Structure of Functional Relationships." *Econometrica* 15 (October 1947): 361–73.

———. "Domestic Production and Foreign Trade: The American Capital Position Re-examined." *Proc. American Philosophical Soc.* 97 (September 1953): 332–49.

———. "Factor Proportions and the Structure of American Trade: Further Theoretical and Empirical Analysis." *Rev. Econ. and Statis.* 38 (November 1956): 386–407.

Meadows, D., et al. *The Limits to Growth.* New York: Universe, 1972.

Meier, R. L. "Information, Resource Use, and Economic Growth." In *Natural Resources and Economic Growth,* edited by J. J. Spengler. Washington: Resources for the Future, 1961.

Moroney, J. R. *The Structure of Production in American Manufacturing.* Chapel Hill: Univ. North Carolina Press, 1972.

Mukerji, V. "A Generalized SMAC Function with Constant Ratios of Elasticities of Substitution." *Rev. Econ. Studies.* 30 (October 1963): 233–36.

Mundlak, Y., and Hoch, I. "Consequences of Alternative Specifications in Estimation of Cobb-Douglas Production Functions." *Econometrica* 33 (October 1965): 814–28.

Office of Business Economics. *Input-Output Structure of the U.S. Economy, 1963.* Vol. 1. *Transactions Data for Detailed Industries.* Washington: Government Printing Office, 1969.

Park, C. F., *Affluence in Jeopardy.* San Francisco: Freeman, Cooper, 1968.

Potter, N., and Christy, F. T., Jr. *Trends in Natural Resource Commodities.* Baltimore: Johns Hopkins Press for Resources for the Future, 1962.

Ruble, W. L. "Improving the Computation of Simultaneous Stochastic Linear Equation Estimates." Ph.D. dissertation, Michigan State Univ. 1968.

Sato, K. "A Two-Level Constant Elasticity of Substitution Production Function." *Rev. Econ. Studies* 34 (August 1967): 201–18.

Schultz, T. W. *The Economic Organization of Agriculture.* New York: McGraw-Hill, 1953.

Shephard, R. W. *Cost and Production Functions.* Princeton, N.J.: Princeton Univ. Press, 1953.

———. *Theory of Cost and Production Functions.* Princeton, N.J.: Princeton Univ. Press, 1970.

Solow, R. M. "The Production Function and the Theory of Capital." *Rev. Econ. Studies* 23 (1955–56): 101–8.

Travis, W. P. *The Theory of Trade and Protection.* Cambridge, Mass.: Harvard Univ. Press, 1964.

U.S. Bureau of the Census. *Census of Manufactures, 1963.* Vol. 1. *Summary and Subject Statistics.* Washington: Government Printing Office, 1966.

———. *Annual Survey of Manufactures: 1964 and 1965.* Washington: Government Printing Office, 1968.

U.S. Bureau of the Census and Board of Governors of the Federal Reserve System. *Census of Mineral Industries, 1963, Indices of Production.* Washington: Government Printing Office, 1970.

Uzawa, H. "Production Functions with Constant Elasticities of Substitution." *Rev. Econ. Studies* 29 (1962): 291–99.

Vanek, J. "The Natural Resource Content of Foreign Trade, 1870–1955, and the Relative Abundance of Natural Resources in the United States." *Rev. Econ. and Statis.* 41 (1959): 146–53.

———. *The Natural Resource Content of United States Foreign Trade, 1870–1955.* Cambridge, Mass.: M.I.T. Press, 1963.

Weiser, L. "Changing Factor Requirements of United States Foreign Trade." *Rev. Econ. and Statis.* 50 (August 1968): 356–60.

Zellner, A. " An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." *J. American Statis. Assoc.* 57 (June 1962): 585–612.

Zellner, A.; Kmenta, J.; and Dreze, J. "Specification and Estimation of Cobb-Douglas Production Function Models." *Econometrica* 34 (October 1966): 784–95.

Capacity Utilization under Alternative Regulatory Restraints: An Analysis of Taxi Markets

---

Arthur S. De Vany

Texas A & M University

Both the Averch-Johnson (A-J) model and the Chamberlin model fail to consider the value of excess capacity to consumers. Service industries, whether they are regulated or not, will usually have excess capacity in the Chamberlinian sense because this capacity conserves time for consumers. This paper examines a model of the taxi industry where allowance is made for capacity to affect the value or quality of the service through its effect on waiting time. The central issue is to determine equilibrium output, capacity, and the utilization of capacity when the market is organized as a franchised monopoly, through a medallion system, and when there is free entry, and to exhibit the relationship among these variables and prices, cost of capacity and output, and policies of the regulator. It is found that many of the characteristics of taxi markets that would appear to confirm the monopolistic-competition thesis arise because of the nature of regulation of these markets.

I. Introduction

The effect of regulation on capacity utilization has been actively debated in recent years. Since the Averch-Johnson demonstration (1962), there has been some acceptance of the idea that regulation designed to secure some rate of return on capital of the firm will induce "excess" capacity. The point also has been made that free entry into markets having some degree of product differentiation results in "excess" capacity.[1] Casual examination of such markets as barbering, real estate sales, stock

I wish to acknowledge the valuable conversations I have had with Richmond Lloyd on the subject matter of this paper. Financial support was provided by the Texas A & M University Organized Research Fund.

[1] Chamberlin (1933) and Kaldor (1935) speak of excess capacity in terms of a rate of production below that which would yield minimum average cost. The A-J result would produce a similar situation with the overcapitalized firm operating where it experienced declining short-run average cost.

[*Journal of Political Economy*, 1975, vol. 83, no. 1]

© 1975 by The University of Chicago. All rights reserved.

brokerage, and taxicabs (in some cities) appears to support this proposition.[2] A recognized deficiency of these demonstrations is the failure to consider the value of capacity to consumers of the product.[3] That is, given any reasonable stochastic properties of demand, capacity serves to reduce waiting time and planning costs for consumers, and, therefore, it is not necessarily inefficient to have some unused capacity. This paper examines a model of an industry where allowance is made for capacity to affect the value or quality of the services received by consumers through its effect on waiting time. The vehicle for this examination is the taxicab industry, but the results obtained are generalizable to a number of service industries where waiting time is likely to be a function of the rate of utilization of each firm's capacity and the number of firms offering the service.

The central issue is to determine the industry response to regulation when the taxi market is organized as a franchised monopoly, when entry is limited through a medallion system, and when there is free entry. My interest is in finding equilibrium output, capacity, and utilization of capacity and in exhibiting the relationship between these variables and prices, the cost of capacity and output, and the policies of the regulator. It is found that a constant cost of quality gives rise to decreasing and then increasing cost of output. The monopolist produces where his marginal cost of output is equal to the regulated price, and this occurs on the nondecreasing portion of the average cost curve, even if profits are zero. Under free entry, each firm produces where marginal cost equals the regulated price at minimum average cost. Monopoly marginal cost is greater than marginal cost under free entry, and so there is less output, less capacity, and lower average cost of output under monopoly. Nonetheless, the monopoly outcome is generally inferior to the free-entry outcome, unless the regulated price is set well above the efficient level.

II. Demand Assumptions

Assume the industry demand function is

$$Q = Q(\pi, \pi_o, B),$$

where Q represents quantity of passenger trips by cab, π is the "full price" of a cab trip, π_o is an index of the full prices of all other goods, and B is a demand-shift parameter. This is the Becker "full price" demand function[4] with the property that $\partial Q/\partial \pi < 0$. Let $\pi = p + vt(H)$, where p is the fare per cab hour, v is the value of time of passengers, and $t(H)$ is waiting time for a cab. The H represents cabs supplied simultaneously and I assume that $t' = \partial t/\partial H < 0$.[5] Differentiating the demand function gives the properties

---

[2] Samuelson (1973, p. 518), for example, suggests barbering is an industry that fulfills the predictions of the Chamberlin model. In this paper, I show that chronic excess capacity may be the result of regulation that prohibits price competition and sets price at an efficiently high level.

[3] Alchian and Allen (1969, pp. 152–57) offer an extensive discussion of the value of excess capacity in the form of empty barber chairs, hotel rooms, retailer inventories, and so on, to consumers having less-than-perfect information and scarce time.

$$\frac{\partial Q}{\partial p} = \frac{\partial Q}{\partial \pi} < 0, \qquad \frac{\partial Q}{\partial H} = \frac{\partial Q}{\partial \pi} vt' > 0.$$

Demand is assumed to be uniformly distributed across the day.

III. Monopoly Solution

Assume a single firm has a charter from a city council or some such agency granting it monopoly rights to pick up passengers within a market area, and that the fare is also set by this agency.[6] It is assumed that cabs can be added to operation at constant cost, so the cost function is

$$C = cH,$$

with c a parameter.[7]

The firm's program is to maximize

$$pQ[p + vt(H)] - cH$$

by choice of H subject to

$$p = p^o,$$
$$Q \leq H,$$

where p^o is the regulated fare. Equilibrium must occur in the interior of the capacity constraint, or else waiting time would be indefinitely long, implying an indefinitely high full price so long as time has a positive value. The solution H^* of

$$p \frac{\partial Q}{\partial H} - c = p \frac{\partial Q}{\partial \pi} vt' - c = 0 \tag{1}$$

---

[4] Becker (1965).

[5] Orr (1969) and Douglas (1972) adopt similar assumptions on the waiting-time distribution, but do not use the full price-demand function.

[6] This description closely characterizes the Los Angeles taxi market (see Eckert 1973).

[7] The cab's cost-minimizing operating period is assumed to span the demand period, and so variable cab hours is strictly equivalent to the number of cabs operated. No distinction is made between cruising or carrying a passenger in the cost function, since it is assumed that the operating cost of a cruising cab does not differ from the cost of a cab engaged in carrying a passenger.

will be a maximum-profit point provided that at that point

$$p\frac{\partial^2 Q}{\partial H^2} = p\left[\frac{\partial Q}{\partial \pi} vt'' + \frac{\partial^2 Q}{\partial \pi^2}(vt')^2\right] < 0. \tag{2}$$

The monopolist operates where marginal revenue of cabs at the regulated price equals marginal cost of cabs, and marginal revenue is declining. By equation (2), these conditions will be fulfilled if the marginal reduction in waiting time due to cabs is declining ($t'' > 0$), and $\partial^2 Q/\partial \pi^2 \leq 0$, or when

$$\frac{\partial^2 Q}{\partial \pi^2} > 0, \quad \text{and} \quad \frac{\partial^2 Q}{\partial \pi^2}(vt')^2 < \left|\frac{\partial Q}{\partial \pi} vt''\right| .$$

Suppose the demand function is of constant elasticity with respect to full price. Now, the price and cab elasticities of demand can be written

$$\beta = \frac{\partial Q}{\partial p}\frac{p}{Q} = \varepsilon\frac{p}{\pi}, \qquad e = \frac{\partial Q}{\partial H}\frac{H}{Q} = \varepsilon vt'\frac{H}{\pi},$$

where ε is the full price elasticity of demand. The consequences of this assumption are that as cab hours are increased, the elasticity of demand with respect to cabs falls, and the price elasticity of demand rises (absolutely). In the neighborhood of equilibrium, quality improvements increase price elasticity of demand and decrease quality elasticity of demand. I shall assume a constant full price elasticity of demand.

Using the definition of elasticity of demand with respect to cab hours, the monopolist's marginal revenue of cabs is

$$MR_H = p\frac{\partial Q}{\partial H} = pe\frac{Q}{H}, \tag{3}$$

and so $MR_H = AR_H$ when $e = 1$, $MR_H > AR_H$ when $e > 1$, and $MR_H < AR_H$ when $e < 1$. The firm operates where $e < 1$ if it earns positive profits; $e = 1$ is a zero-profit point, and $e > 1$ is a negative-profit point. The monopolist operates where the occupancy rate of its cabs declines with further increases in its supply of cabs.

Equilibrium is depicted diagrammatically in figure 1. The long-run average and marginal cost curve of cab hours is $AC_H = MC_H$, and marginal and average revenue of cab hours are MR_H and AR_H at the regulated price. Equilibrium hours are H^m.

These relationships are particularly interesting when translated from cab hours into passenger trips. What is revealing about this translation is that the supply of cab hours H may now be regarded as a dimension of quality of output. Consider first the relationship between H and the demand curve. An increase in H reduces time cost and increases the quantity demanded at each price—the extent of this increase depends on

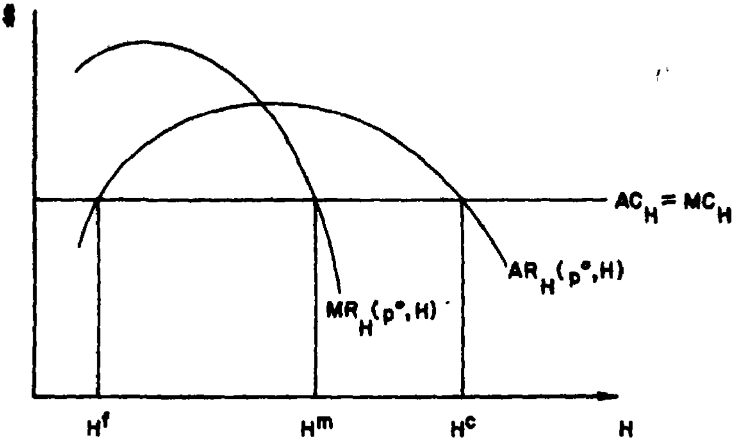

FIG. 1.—Equilibrium cab hours at p^*

the value of time, the reduction in waiting time achieved, and the full price elasticity of demand. Under the assumptions behind the AR_H curve, the demand curve shifts more than proportionately with increases in H initially, and then less than proportionately as ϵ falls below unity, with the firm settling into the latter region.

Now consider how costs expressed in terms of passenger output behave as H is varied. Define the output-cost function $C(Q)$ by

$$C(Q) - cH \equiv 0. \tag{4}$$

Using the implicit-function rule gives

$$\frac{\partial C}{\partial Q} = \frac{c}{\partial Q/\partial H} \tag{5}$$

as marginal cost of output MC_Q. Marginal cost is positive since $\partial Q/\partial H > 0$, and rising at equilibrium, since $\partial^2 Q/\partial H^2 < 0$ by equation (2). Average cost of output is

$$AC_Q = \frac{C(Q)}{Q} = \frac{cH}{Q},$$

and, using the definition of ϵ in (5),

$$MC_Q = \frac{cH}{Q\epsilon}. \tag{6}$$

So $MC_Q = AC_Q$ where $\epsilon = 1$, and at this point AC_Q is at a minimum, while AR_H is at a maximum. Thus, the output-cost curves are U-shaped, even though cost per cab hour is constant. These properties are reflected in the cost curves drawn in figure 2. Furthermore, applying equation (3)

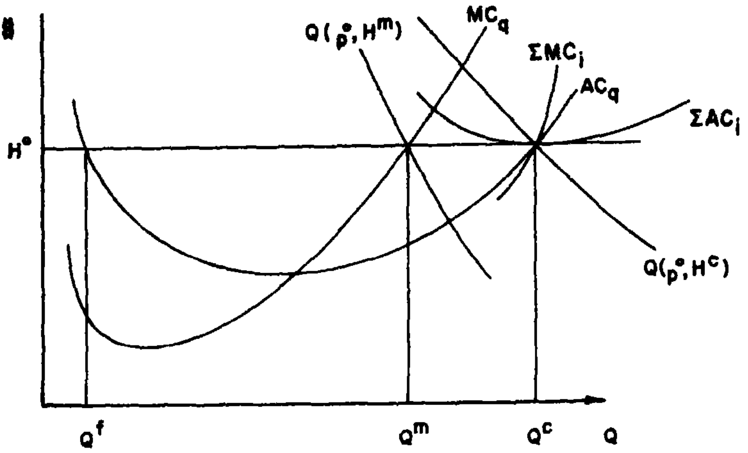

Fig. 2.—Equilibrium output at p^o

to equation (1) gives

$$pe\frac{Q}{H} - c = 0,$$

so

$$p = \frac{cH}{Qe}, \tag{7}$$

from which it follows that, in equilibrium, marginal cost of output is equal to the regulated price.[8] It should be noted that, even though the firm produces with excess capacity, it produces to the right of the minimum point of the average cost curve as long as it is earning profits. A zero-profit equilibrium occurs at minimum average cost and not at a point of tangency of the average cost-and-demand curves.

IV. Competitive Solution

Assume the industry is comprised of owner-operated cabs, one cab per owner or firm. The price to be charged is set by the regulator, but entry is unrestricted. If there are any economies associated with maintenance of large fleets and in the dispatching of cabs, these services can be provided by an independent maintenance contractor and by an impartial dispatcher maintained by an association of the cab owners. Assume operating cost per cab hour is U-shaped, declining initially because of some fixed costs, and eventually rising as the cab is operated ever more hours per day because of mounting maintenance costs and because costs include the imputed value of the owner's time, which rises as he drives more hours per day, leaving less time for other uses. Assume cabs are operated at the rate that minimizes cost per cab hour and that this cost is c. Assume also that this operating period spans the demand period. Then new cabs may be added at constant cost, and industry cost will be cH, where $H = \sum h_i$.

---

[8] Since it is fixed, price is marginal revenue, so this is not an unexpected result. The novelty is in the fact that, at the going price, output may only be varied by varying product quality, and it is the cost of changing output in this manner that counts.

Given a reasonably large number of cabs, there will be no differentiation of cabs: the information structure assumed is that each potential cab rider knows that there is some expected interval between wanting a cab and getting one and uses this knowledge to form expected price. Aggregating individual demand functions, since price is fixed and equal for all cabs, gives the industry demand function

$$Q = Q[p + vt(H)].$$

Industry equilibrium occurs where average revenue per cab hour equals average cost; this is point H^c in figure 1. Under the cost assumptions, both the franchised monopoly and the competitive industry will have identical, and constant, costs with respect to cab hours.

Assuming market share equals share of industry cab hours, the profit function of the individual firm is

$$pQ[p + vt(H)]\,\frac{h_i}{H} - C(h). \tag{8}$$

The ith firm's program is to maximize profits by choice of h_i subject to the regulated fare. The maximum-profit point satisfies

$$p\,\frac{\partial q_i}{\partial h_i} - \frac{\partial C}{\partial h_i} = 0, \tag{9}$$

which, using the definition $e_i = \partial q_i/\partial h_i(h_i/q_i)$, may be written

$$pe_i - \frac{\partial C}{\partial h}\frac{h_i}{q_i} = 0. \tag{10}$$

The elasticity of demand with respect to cab hours experienced by the firm may be written

$$e_i = 1 + (e - 1)\,\frac{h_i}{H} + (e - 1)\,\frac{h_i}{H}\sum_{j=1}^{n-1}\frac{\partial h_j}{\partial h_i},$$

where $\partial h_j/\partial h_i$ is the change in rival's cab hours induced by i's change in cab hours, and e is the industry elasticity. If it is assumed that there are a large number of competing cab companies, each operating independently of the others, then $h_i/H \rightarrow 0$, so that $e_i \rightarrow 1$. This means that the firm will feel that it can sell all the cab hours it wishes to supply at the industry equilibrium-utilization rate. When $e_t = 1$, then from equation (10)

$$p - \frac{\partial C}{\partial h}\frac{h_t}{q_t} = 0, \tag{11}$$

and price is equal to the marginal cost of capacity (cab hours) times capacity per unit of output. In terms of output, rather than cab hours, we have, as in the monopoly case

$$p = \frac{\partial C}{\partial h}\frac{h_t}{q_t e_t}. \tag{12}$$

But, since $e_t = 1$,

$$p = \frac{\partial C}{\partial h}\frac{h_t}{q_t} \tag{13}$$

for each firm, and price is equal to marginal cost of output. Since price also equals average cost, marginal and average cost will be equal, implying that output is produced at minimum average cost for the equilibrium level of capacity utilization.

The industry marginal- and average-cost curves are U-shaped and identical to the franchised-monopoly curves. As entry initially occurs, waiting time falls more than proportionately, so that demand increases by more than the supply of cab hours, the utilization rate increases, and average cost per unit of Q falls. In this initial stage, entry of new firms increases profits of existing firms, since the average utilization rate realized by existing firms rises when new firms enter. This is not a stable situation, and it is clear that industry equilibrium may only occur in the nondecreasing range of the industry cost curve, where further entry decreases profits of existing firms. The competitive industry equilibrium occurs at Q^c in figure 2. The competitive output is greater than monopoly output, the competitive industry demand curve lies above the monopoly demand curve—because of the higher quality or lower waiting time—and average cost at the competitive output is higher than at the monopoly output, though marginal cost is equal for both.

The industry marginal-cost-of-output curve is also the monopolist's marginal cost curve, but it is not the sum of the MC curves of the individual firms. The reason for this is that the industry, on the whole, experiences decreasing returns to cab hours, but each individual firm realizes constant returns at any point of industry equilibrium. The sums of the marginal- and average-cost curves of all firms are shown as $\sum MC$ and $\sum AC$ at the industry equilibrium point. The average cost curve shows social marginal cost, but only at the regulated price, it is not the long-run industry supply curve.

V. The Medallion Solution

Assume one medallion per cab, and, as before, that the optimal operating period of the cab spans the demand period. Referring to figure 1, the feasible region begins at H^f, where revenue covers the cost of providing the service. If the number of medallions issued were less than H^f, none would be taken and no service would be provided. Any number of medallions in excess of H^c would not be taken if there were free entry, and no more than H^m would be taken by the monopoly. The rent per medallion is the difference between *AC* and *AR* associated with the number in existence. The medallion solution ranges between H^f and H^c, with zero rents and no service at H^f, and zero rents and maximum service at H^c. The effect of increasing the number of medallions on rents depends upon the initial conditions: if initially the number of medallions is less than H^m, then rents will increase as more are issued, but will fall with more medallions to the right of H^m.

VI. Limited Entry with Unrestrained Price

Suppose now the regulator simply sets the number of cabs allowed to operate, but leaves price free to find its own equilibrium. There is insufficient space in this paper to consider this problem in a thorough manner; I confine myself to a description of equilibrium when it is assumed that there are a large number of individually operated cabs competing for sufficiently well-informed passengers such that price equals average cost. Then price is a well-defined function of the number of cabs and the waiting-time function.

From the assumption of constant industry cost in cab hours, it follows that the long-run average cost of output is a U-shaped curve, because of the increasing and then decreasing productivity of cab hours in reducing delay. The *LRAC* curve is drawn in figure 3. The open-entry competitive equilibrium will be on the nondecreasing portion of this curve; assume it is point Q^o. If the regulator limits cabs to less than the quantity associated with Q^o, there will be longer waiting time, lower effective demand, less output, and a lower price. It can be shown that Q^o is the efficient level of output for the industry if it is the open competitive equilibrium.[9] If the regulator imposed Q' as a limit (through imposition of a limit on cabs), average cost and price would be lower than at Q^o, but the demand curve would also be shifted back from D^o to D', thus reducing consumer surplus. The result of this limitation would be a reduction in welfare, because the reduction in price achieved is too little to offset the reduction in service quality—that is, the increase in waiting time.[10] On the other hand, if the number of cabs were severely enough limited

---

[9] This is proved in De Vany (1973).

[10] This would follow if Q^o were efficient as asserted.

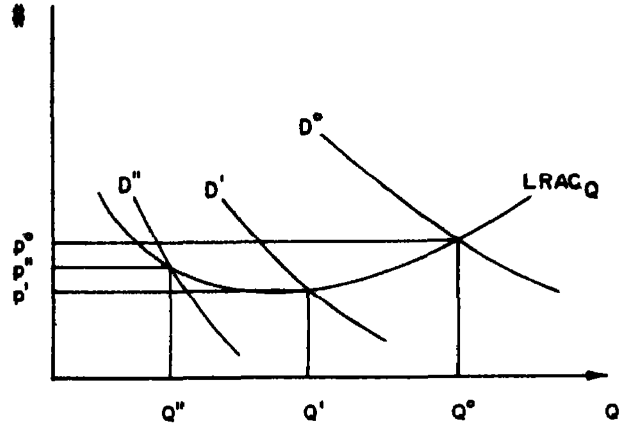

FIG. 3.—Equilibrium with entry limitations and unregulated price

so as to result in, say, output Q'', the demand curve would be shifted back to D'', and price and average cost would be increased. In this region it would appear that the industry experiences decreasing costs.

VII. Efficient Pricing

Assuming equal value of time for all consumers, the efficient levels of output and cab hours (or waiting time) occur where

$$-vt' = c,$$

$$-vt' = p\,\frac{Q}{H}, \tag{14}$$

or, equivalently,

$$p = c\,\frac{H}{Q}. \tag{15}$$

The first condition requires that the marginal value of capacity equal its marginal cost, and the second that average and marginal value of capacity be equal. Together, these conditions give equation (15), which indicates that price is equal to marginal cost of output, or, what is the same thing, price equals marginal cost of capacity times the number of units of capacity per unit of output.[11]

These conditions are equivalent to those that ensure that Q is maximized subject to a zero-profit constraint, and so the efficient price maximizes output.[12] The effect of price variation on monopoly supply of cab hours is:

---

[11] Since it is assumed that the marginal cost of carrying a passenger on a cab that is already cruising is zero, to set price equal to this marginal cost would create losses for the industry. Efficient pricing demands that price equal marginal cost of output at the efficient level of capacity. True marginal cost is the cost of adding a passenger without reducing quality, i.e., holding H/Q constant. This is given by cH/Q, which equals price in competitive equilibrium. Cost is rising at equilibrium, since $\epsilon > 0$; and industry marginal cost $cH/Q\epsilon$ exceeds average cost, since $\epsilon < 1$. Monopoly marginal cost $cH/Q\epsilon$ is greater than true marginal cost, so monopoly is inefficient, even though price equals marginal cost at monopoly equilibrium.

$$\frac{\partial H}{\partial p} = -\frac{\beta + 1 + \varepsilon_p^e}{p\,\partial^2 Q/\partial H^2}, \tag{16}$$

where $\varepsilon_p^e \leq 0$ is the elasticity of the capacity elasticity of demand with respect to price.[13] A price increase will increase monopoly capacity only if demand is quite price inelastic. In the case of open entry, the change in industry capacity is

$$\frac{\partial H}{\partial p} = -\frac{Q(\beta + 1)}{pQe - cH}, \tag{17}$$

which, since $(pQe - cH) < 0$, depends upon the price elasticity of demand as well. The competitive derivative lacks ε_p^e, and so a comparable increase in the regulated price will be more likely to expand capacity under competition than under monopoly. For either case, the change in output due to a change in the regulated price is

$$\frac{dQ}{dp} = \frac{\partial Q}{\partial \pi} + \frac{\partial Q}{\partial \pi} v \frac{\partial t}{\partial H} \frac{\partial H}{\partial p}, \tag{18}$$

and the full change in output depends upon a (negative) direct price effect and an indirect induced quality effect which may be positive or negative.

The efficient, or output-maximizing, price occurs where (18) is zero. Since $\partial H/\partial \pi$ and $\partial t/\partial H$ are negative, $\partial H/\partial p$ must be positive at maximum output, implying that output is maximized at a lower price than capacity. The price which maximizes monopoly output is less than the efficient price which maximizes competitive output, and monopoly output and capacity are always less than competitive output and capacity. These relationships are shown in figure 4. A price higher than the efficient one (p^*) will reduce competitive output and expand capacity beyond the efficient level, as long as price is not set higher than the level that maximizes capacity (p'). A higher price than the capacity-maximizing one reduces both competitive capacity and output. The competitive outcome is superior to the monopoly outcome at every price up to and including the efficient level. At a price above p^*, competitive capacity will be expanded relative to output, and the marginal cost of capacity will exceed its marginal value. There will exist a price above the efficient level at which marginal cost of capacity under competition will be so far

---

[12] Douglas (1972) shows this to be true. However, the present model makes no allowance for such possible externalities as street congestion caused by taxis. Nor does it allow for differences in the value of time of consumers, in which case a uniform price would prevent product differentiation.

[13] Implicit differentiation of eq. (1) and some manipulation yields this result.

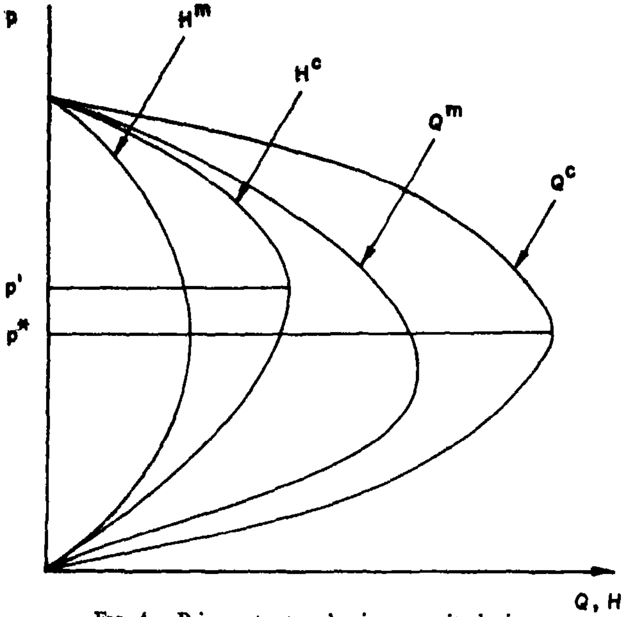

FIG. 4.—Price-output and price-capacity loci

in excess of marginal value that the lesser capacity offering of the monopolist will be more efficient than the competitively supplied capacity. This does not come about because competition is wasteful; rather, it is caused by the regulator's setting price above the efficient level.

References

Alchian, A. A., and Allen, W. R. *Exchange and Production Theory in Use.* Belmont, Calif.: Wadsworth, 1969.

Averch, H., and Johnson, C. C. "Behavior of the Firm under Regulatory Constraint." *A.E.R.* 52 (December 1962): 1053–69.

Becker, Gary S. "A Theory of the Allocation of Time." *Econ. J.*, vol. 75 (September 1965).

Chamberlin, E. H. *The Theory of Monopolistic Competition.* Cambridge, Mass.: Harvard Univ. Press, 1933.

De Vany, A. "The Allocation of Time in Competitive Markets." Mimeographed. Department of Economics, Texas A & M Univ., 1973.

Douglas, G. W. "Price Regulation and Optimal Service Standards: The Taxicab Industry." *J. Transport Econ. and Policy* 20 (May 1972): 116–127.

Eckert, R. D. "On the Incentives of Regulators: The Case of Taxicabs." *Public Choice* 14 (Spring 1973): 83–101.

Friedman, M. *Price Theory.* Chicago: Aldine, 1962.

Kaldor, N. "Market Imperfection and Excess Capacity." *Economica* 11, n.s. (1935): 33–50.

Orr, D. "The Taxicab Problem: A Proposed Solution." *J.P.E.* 77, no. 1 (January/February 1969): 141–47.

Samuelson, P. A. *Economics.* 9th ed. New York: McGraw-Hill, 1973.

An Estimate of the Liquidity Premium

J. Huston McCulloch

Boston College and Harvard University

The liquidity premium on U.S. government securities is quantitatively estimated and tabulated, using maturities from 1 month to 30 years. Unbiased forecasting by the market is assumed in order to get at expectations. The premium is estimated, first allowing it to take any shape and then constraining it to conform to a functional form which implies that the "normal" shape of the yield curve is monotonically increasing toward an asymptote. Tests for constancy of the premium over the post-Accord period, normality of the forecasting errors, and monotonicity of the premium with respect to maturity are performed, and the dependence of the premium on the level of interest rates is discussed.

I. Introduction

It has long been known that short-term interest rates tend to lie below long-term rates on securities with equal default risk. As early as 1935, Charles C. Abbott referred to the reversals in the usual relationship that occurred from June 1920 to January 1921 and from May 1928 to November 1929 as "the more striking in that they contradict the well-known tendency for obligations of short maturity to sell consistently at higher prices and with lower yields than obligations of equal security but longer maturity."[1]

This paper is based on the author's Ph.D. dissertation, Economics Department, University of Chicago. He is indebted to Lester G. Telser, the chairman of his thesis committee, for encouragement and guidance. He also benefited from discussions with, and suggestions from, the other members of his committee, Reuben A. Kessel and Merton H. Miller. In addition, he wishes to acknowledge helpful comments at various stages in this research from Fischer Black, Eugene F. Fama, James Meginnis, Charles R. Nelson, members of the Applied Price Theory and Money and Banking Workshops at the University of Chicago, and the referees. The author was supported by a fellowship from the Earhart Foundation.

[1] Abbott (1935, p. 9). As it happens, the evidence Abbott cited was not statistically strong enough to demonstrate the existence of a liquidity premium (McCulloch 1973, pp. 62–63).

[*Journal of Political Economy*, 1975, vol. 83, no. 1]

© 1975 by The University of Chicago. All rights reserved.

A number of questions regarding this "liquidity premium" need to be answered. How large is it for various pairs of maturities? Does the premium increase monotonically with maturity? Does it vary with the level of interest rates? Is it the same for different periods in time? And is it large enough to imply that holders of longer-term bonds have a systematically higher holding period yield, after taking into account transactions costs as reflected by bid-asked spreads?

The liquidity premium is defined as the difference between a forward interest rate and the market's expectation of the corresponding future spot rate. Measurement of this premium is complicated by the fact that expectations are not directly observable. A straightforward, but in practice unreliable and limited, method is to ask market participants what they believe future interest rates will be (Kane and Malkiel 1967). Another approach is to build a model of expectation generation based on the past behavior of interest rates and, in some instances, of other variables.[2] However, these models entail strong assumptions about the set of information people take into account in forming their expectations. A third approach simply assumes that forecasting errors have mean zero, so that the subsequent realization gives us an unbiased estimator of the forecast. In one variation on this third approach, Kessel (1965) actually compares the forward rate to the subsequent spot rate. However, this variation does not make the most efficient use of the data available. Under the same basic assumption, Cagan (1969) and Roll (1970) have been able to use the term structure for two nearby points in time to obtain observations on the liquidity premium all the way out to the longest maturity observed.

Unfortunately, neither Cagan's paper nor Roll's book tells us about the size of the liquidity premium applicable to forecasts for periods longer than a few months. For instance, at the end of June 1953, the forward rate on a 10-year loan to begin 5 years in the future was 3.16 (±0.01) percent per year. Five years later, the corresponding 10-year spot rate was 2.94 (±0.02) percent per year.[3] It would be interesting to know whether such a fall was unanticipated, whether it is about the size one would expect given liquidity preference, or whether the difference was less than the mean liquidity premium, so that it actually represents an unanticipated rise. The answer has not been given in the literature for maturities of this duration.

In this paper we use a modification of the Cagan-Roll method to estimate the size of the post-Accord liquidity premium for all available maturities, together with the relevant measurement errors implied by this procedure, and to attack the other questions posed above. The results are compared with those obtained by other investigators.

---

[2] E.g., Meiselman (1962), Diller (1969), Nelson (1972), Modigliani and Shiller (1973), Modigliani and Sutch (1966).

[3] Calculated by the author, using bid-asked mean prices of U.S. government securities and the technique described in McCulloch (1971).

II. The Behavior of the Postwar Liquidity Premium

In this section, we define an estimator of the liquidity premium. Its values for three pilot maturities are used to test the premium and the variance of the forecasting errors for constancy over the postwar period. The forecasting errors are tested for normality, and the dependence of the liquidity premium on the level of interest rates is discussed.

An Estimator of the Liquidity Premium

In the real world, observed bond maturities do not lie at evenly spaced discrete intervals, even though this would make our calculations easier. Nevertheless, we would expect the price of any security to be governed by a smooth discount function $\delta(t, s)$, which gives the value at time t of a dollar to be repaid at time s in the future, that is, after maturity $m = s - t$.

Corresponding to the discount function we observe at time t is a two-dimensional complex of forward interest rates $r(t, s_1, s_2)$ on hypothetical point-payment forward loans to begin at time s_1 and be repaid at time s_2, where $t \leq s_1 \leq s_2$. Most past investigators have taken m_2 (where $m_2 = s_2 - s_1$, the duration of the forward loan) as a constant with some convenient value, usually a week, month, quarter, or year, and have considered the liquidity premium corresponding only to a one-dimensional complex of forward rates as m_1 varies (where $m_1 = s_1 - t$, the period until the forward loan begins). We, on the other hand, are interested in all values of m_2, from the limit as m_2 approaches zero out to several years or even decades. In the next section, we will treat both forward rates and the liquidity premium in terms of the two variables m_1 and m_2.

However, the forward rates for different values of m_2 are far from independent. In general, forward rates with large m_2 can be obtained by averaging together appropriate forward rates with smaller m_2. It is therefore convenient for our investigation of the qualitative properties of the liquidity premium to single out a one-dimensional complex of independent forward rates, taking m_2 equal to its smallest value of interest. In our case, this is the limit as m_2 goes to zero. This "marginal" or "instantaneous" forward rate $\rho(t, s)$, where $s = t + m_1$, is related to the discount function by

$$\rho(t, s) = -100 \frac{\partial\delta(t, s)/\partial s}{\delta(t, s)}. \tag{1}$$

Because we fit a smooth curve to the discount function, we are able to evaluate this derivative. When t and s coincide, $\rho(t, t)$ becomes a "spot" rate of interest on a hypothetical loan of very short maturity, in effect a "call money" rate.

We define the liquidity premium $\pi(m)$ to be the difference between the forward rate and the expected value of the future spot rate:

$$\rho(t, s) = E_t\rho(s, s) + \pi(m), \qquad m = s - t, \tag{2}$$

where E_t denotes the expected value as of time t.[4] Of course, not all participants have the same expectations about the future. We must regard $E_t\rho(t, s)$ as some sort of market average of individual expectations.[5]

If forecasting errors are unbiased, we could compare the forward rate with the actual subsequent spot rate to get an estimator of $\pi(m)$. However, for large values of m, we would have very few such pairs of observations. We can get around this problem by observing that today's forecast of some distant future variable must be an unbiased estimator of all future forecasts of that variable. Thus, if we have observations on the term structure at points in time Δt apart, we must have:

$$E_t\rho(s, s) = E_t[E_{t+\Delta t}\rho(s, s)], \tag{3}$$

where s is some time later than $t + \Delta t$. This seems like a reasonable assertion; if we had any reason at time t to believe that our next period expectation of $\rho(s, s)$ would be any different from our current expectation, we would already have incorporated this information into our current expectation, bringing the two expectations into equality.[6]

Given observations on the term structure at times $t_1, t_2, \ldots t_j \ldots,$ where $t_j - t_{j-1} = \Delta t$, we define $\Delta\pi_j(i\Delta t)$ by

$$\Delta\pi_j(i\Delta t) = \rho(t_j, t_j + i\Delta t) - \rho(t_j + \Delta t, t_j + i\Delta t). \tag{4}$$

Putting (2), (3), and (4) together, we have

$$\Delta\pi_j(i\Delta t) = \pi(i\Delta t) - \pi[(i - 1)\Delta t] + u_{i,j}, \tag{5}$$

---

[4] Two slightly different possible definitions of the liquidity premium are discussed in McCulloch (1973, pp. 18–19).

[5] See Williams (1938), chap. 10. Bierwag and Grove (1967) develop an ingenious model in which the market behaves as if governed by a market expectation which is an arithmetic average of individual expectations, weighted in proportion to the individuals' audacity (negative risk aversion), certainty, and wealth. Unfortunately, they deal in terms of price uncertainty, instead of consumption uncertainty. Compare Stiglitz (1970), who correctly deals with consumption uncertainty, although he assumes homogeneous expectations.

[6] This is essentially the point Samuelson makes (1965, pp. 41–49). His analysis is unnecessarily complicated by assumptions about the relation between anticipations and current forward prices. The real crux of his argument is that anticipations themselves must fluctuate randomly. The proof of our assertion is similar to the proofs he gives. Once we have identified the psychological anticipation with the mathematical expected value by our unbiased forecasting assumption, the remainder of the proof is a purely mathematical consequence of the properties of probability distributions.

where $u_{i,j}$ is a random forecasting error with mean zero. When we sum (5) over i from one to n, we obtain the following estimator of $\pi(m)$ for $m = n\Delta t$:

$$\begin{aligned}\pi_j(m) &= \sum_{i=1}^{n} \Delta\pi_j(i\Delta t),\\ &= \pi(m) - \pi(0) + \sum_{i=1}^{n} u_{i,j}, \qquad (6)\\ &= \pi(m) + v_{m,j}.\end{aligned}$$

Since the $u_{i,j}$ have mean zero, the $v_{m,j}$ also have mean zero.[7]

The discount function $\delta(t, s)$ was fit for the close of each month from December 1946 to March 1966 by means of a quadratic spline.[8] The prices used were the means of bid-and-asked monthly closing offers for most fully taxable U.S. government bills, notes, and bonds.[9] Forward rates were derived from these discount curves, and from them the values $\pi_j(m)$ were calculated. Figure 1 shows monthly observations on $\pi_j(m)$ for $m = 1$ year, along with the average yield to maturity

$$\begin{aligned}\eta(t_j, m) &= \frac{1}{m}\int_{t_j}^{t_j+m} \rho(t_j, s)\, ds \qquad (7)\\ &= -\frac{100}{m}\ln \delta(t_j, t_j + m).\end{aligned}$$

For ease of comparison, both are shown relative to the same origin, but with different scales.

Likelihood Ratio Tests for Homogeneity

For the purpose of analysis, we divided this interval into four approximately equal periods, as shown in table 1. These periods correspond roughly to the pre-Accord period, the first Eisenhower administration,

---

[7] Roll (1970, pp. 98–99) estimates the liquidity premium as the sum of first differences of forward rates, so this is basically the same method as his. However, he used forward rates spanning 1 week, while we have taken the limit as the period spanned goes to zero.

[8] See McCulloch (1971) for details of this procedure, and Rice (1969, 2:123–67) for a discussion of splines in general. Compare Williams (1938, pp. 120–24), who gives an algorithm that fits security prices exactly. His method yields a step-function forward curve (see his chart 4, p. 355), instead of one that is continuous. The method proposed by Bryan and Carleton (1972) is similar to his in this respect. Weingartner (1966) proposes a method based on successive approximations to the coupon-free yield curve. Buse (1970) raises objections that apply to a large number of other studies.

[9] These data were collected by Reuben A. Kessel from the quotation sheets of Salomon Brothers and Hutzler, Inc., of New York, and were processed under the supervision of Merton H. Miller and Myron Scholes. Reduction of these data into discount curves and forward rates was supported by a grant from the University of Chicago, Graduate School of Business.

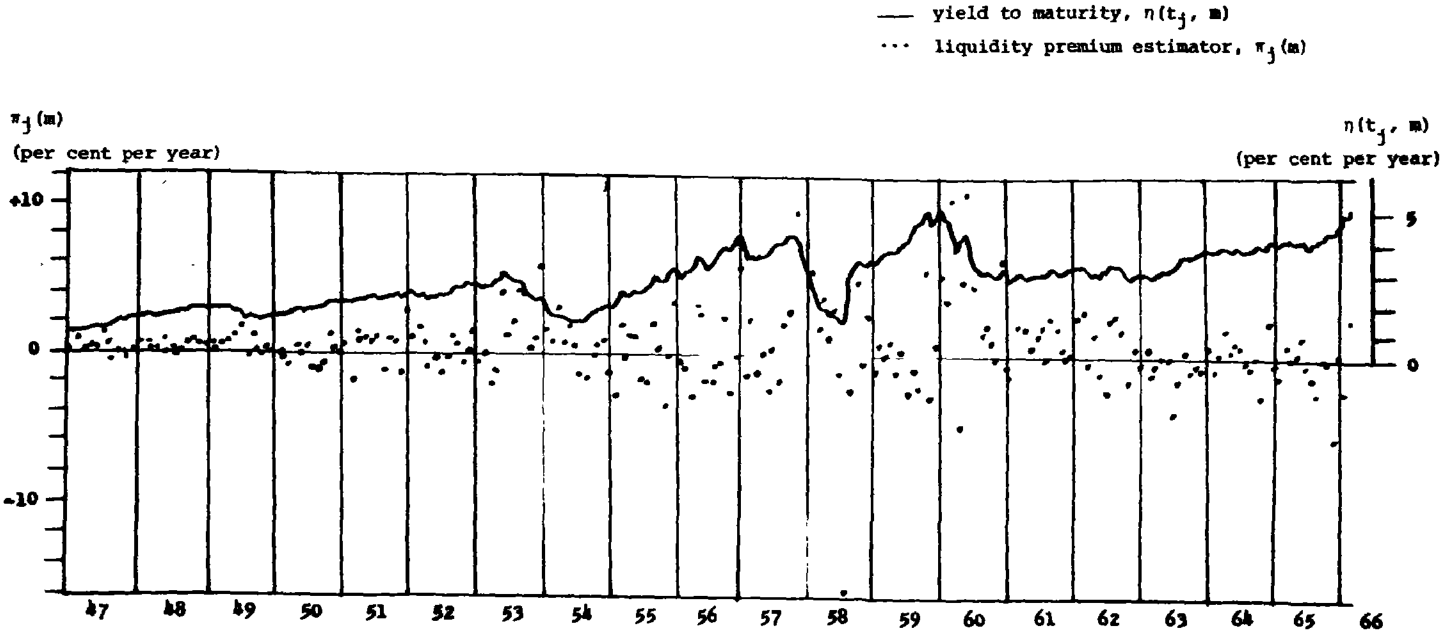

FIG. 1.—Solid line-yield to maturity, $\eta(t_j, m)$; dotted line-liquidity premium estimator, $\pi_j(m)$

TABLE 1

Definition of Periods for Homogeneity Tests

Period	Number of Even-numbered Observations
1. December 31, 1946–March 4, 1951	25
2. March 4, 1951–December 31, 1955	29
3. December 31, 1955–December 31, 1960...........	30
4. December 31, 1960–March 31, 1966	31

the second Eisenhower administration, and the Kennedy/Johnson administration.

Although we would expect the pure forecasting errors to be serially uncorrelated, the errors we observe may not be because of additional errors introduced in measuring $\pi_j(m)$. The term structure cannot be measured exactly, and measurement errors for forward rates of different maturities are strongly correlated. Adjacent estimators $\pi_{j-1}(m)$ and $\pi_j(m)$ are both dependent on the term structure for time t_j, so their measurement errors will not be independent. However, the measurement errors for alternate estimators are completely independent, so that if we discard alternate observations, we should satisfy the serial independence assumption necessary to estimate the mean. Using all the data would give us smaller, but erroneous, confidence intervals. In order to avoid the month of the Accord, we will use the even-numbered observations, that is, starting with the month January 31, 1947 to February 29, 1947.[10]

The premium seems to be small with a small variance during period 1, definitely positive with an intermediate variance during periods 2 and 4, and perhaps larger with the largest variance during period 3. This suggests the following hypotheses: H_1—a common mean and variance for all four periods (one mean, one variance); H_2—one mean and variance for period 1, and a second mean and variance for periods 2–4 (two means, two variances); H_3—one mean for all four periods, but separate variances for each period (one mean, four variances); H_4—one mean for period 1, a second mean for periods 2–4, and separate variances for each period (two means, four variances); H_5—a different mean and variance for each period (four means, four variances).

These hypotheses were compared to one another using the likelihood ratio test. The asymptotically χ^2 statistics are shown in table 2, along with the relevant critical values of the appropriate χ^2 distribution.

We see that we may easily reject H_1 in favor of H_2, and H_2 in favor of H_4. We may reject H_3 in favor of H_4 for $m = 3$ months but get insignificant values for this test for 1 year and 15 years. We may not,

---

[10] It will be shown below, in table 3, that while all the observations taken together exhibit autocorrelation, alternate ones do not.

TABLE 2

LIKELIHOOD RATIO χ^2 STATISTICS

HYPOTHESIS		MATURITY			DEGREES OF FREEDOM	CRITICAL LEVELS OF χ^2				
						Degrees of Freedom	Confidence Level			
Null	Alternative	3-month	1-year	15-year			.995	.99	.90	.80
H_1	H_2	48.3	53.6	34.0	2	1	7.9	6.6	2.7	1.6
H_2	H_4	51.9	44.2	19.8	2	2	10.6	9.2	4.6	3.2
H_3	H_4	14.7	1.2	0.0	1					
H_4	H_5	3.8	1.2	4.4	2					

NOTE.—Based on 115 even observations.

TABLE 3

MAXIMUM-LIKELIHOOD ESTIMATES OF MEANS AND STANDARD DEVIATIONS UNDER H_4

			MATURITY		
STATISTIC	PERIOD	OBSERVATIONS	3 months	1 year	15 years
$\hat{\mu}$	1	Evens	0.052 (0.046)	0.06 (0.12)	−0.8 (1.2)
$\hat{\mu}$	2–4	Evens	0.357 (0.057)	0.33 (0.22)	−0.7 (1.7)
$\hat{\mu}$	2–4	Odds	0.394 (0.054)	0.50 (0.22)	0.1 (1.5)
$\hat{\sigma}$	1	Evens	0.23	0.61	5.88
$\hat{\sigma}$	2	Evens	0.58	1.67	14.03
$\hat{\sigma}$	3	Evens	1.42	4.67	27.06
$\hat{\sigma}$	4	Evens	0.38	1.68	13.25
vNR	1–4	All (231 observations)	1.33 (0.999999+)	1.67 (0.989)	1.90 (0.58)
vNR	1–4	Evens (115 observations)	1.82 (0.71)	1.82 (0.70)	1.77 (0.82)

NOTE.—Means "μ" are in percent per year. Standard deviations "σ" are in (percent per year) per $(\text{month})^{1/2}$. Confidence levels shown in parentheses beneath the von Neumann ratios (vNR) are for a two-tailed test against no autocorrelation, based on the normal approximation for large samples given in Hart (1942).

however, reject H_4 in favor of H_5. We interpret these results as meaning that there has been a fairly constant liquidity premium since the Accord, although the variance of forecasting errors has not been constant. We may not treat the pre-Accord liquidity premium as equal to the post-Accord premium, but this is to be expected, since prior to the Accord the Federal Reserve System explicitly supported the prices of government securities, so that forward rates did not necessarily reflect market forces. The difference between the pre- and post-Accord means for 1 year and 15 years is not significant due to the greater accumulation of forecasting errors for these maturities, but we assume that the same is true for these maturities as for 3 months.[11]

Table 3 shows the maximum-likelihood estimates of the means and standard deviations for the different periods under H_4. The pre-Accord premium, at least for 3 months, was significantly lower than the post-Accord premium. Standard deviations increase with maturity and are higher for period 3 and lower for period 1 than they are for periods 2 or 4. For comparison, the post-Accord means for the omitted odd observations are shown. They run a little higher than the means for the even observations, but not significantly so. Also shown are von Neumann ratios based on the H_4 residuals divided by their estimated standard deviations. As predicted above, autocorrelation is highly significant when we use all the observations (at least for 3 months and 1 year), but becomes insignificant when we use only alternative observations.

---

[11] Wallace (1964, pp. 25–26) also finds significantly different behavior prior to the accord than after.

TABLE 4

STANDARDIZED RANGE TESTS FOR NORMALITY

	3 months	1 year	15 years
Raw data:			
Standardized range	9.48	8.85	6.41
$\hat{\alpha}_{.96}$	1.34	1.47	1.47
After adjustment for heteroskedasticity:			
Standardized range	5.38	5.74	5.53

Confidence Intervals for Normality—90 Observations*
(Two-tailed Test)

.80	4.36–5.60
.90	4.23–5.82
.98	4.01–6.27
.99	3.95–6.44

NOTE.—Based on 90 even post-Accord observations.
* Interpolated from David, Hartley, and Pearson (1954).

Tests for Normality

Table 4 shows the "standardized range" statistic for our post-Accord observations $\pi_j(m)$. It is defined as the ratio of the sample range to the estimate of the standard deviation of the errors. When the standard deviation is treated as if it were constant, we find that we may easily reject normality at the 99 percent level for 3 months and 1 year, and at the 98 percent level for 15 years. This finding might lead us to reject, as Roll does (1970, chap. 4), normality in favor of the class of symmetric stable distributions. If so, we would arrive at the estimates for the characteristic exponent α of the stable distributions shown in the second line of table 4.[12] Our estimates of α are on the same order as those obtained by Roll.[13] However, when we allow for different variances in periods 2, 3, and 4, and divide the maximum-likelihood residuals by their respective estimated standard deviations before calculating the standardized range, we find that the resulting statistics are never significant at the 90 percent level, and only once at the 80 percent level. Although we can reject homoskedastic normality, we are unable to reject heteroskedastic normality. Provided we allow for this heteroskedasticity, we seem to be justified in using tests based on the more familiar normal distribution.[14]

[12] See Fama and Roll (1968; 1971) for techniques of estimating α. The estimator $\hat{\alpha}_{.96}$, which is used here, is based on the .04 and .96 fractiles of the sample distribution. Our "standardized range" is the same as Fama and Roll's "Studentized range."

[13] Roll finds values of $\hat{\alpha}$ from 1.22 to 1.72, mostly around 1.4, for 1949–64 weekly data. For 17–22 weeks to maturity, he finds some as low as 1.00 for 1959–64 (1970, p. 70). A Cauchy distribution has $\alpha = 1$, while a normal distribution has $\alpha = 2$.

[14] Press (1967) and Praetz (1972) also offer interpretations of leptokurtic residuals in terms of mixtures of normals with different variances. However, the compound distributions they suggest are homogeneous over time, in contrast with the one we propose.

Dependence of the Liquidity Premium on the Level of Interest Rates

Van Horne (1965) and Nelson (1972, p. 93) offer evidence which, they contend, supports the view that the liquidity premium decreases with the level of interest rates. On the other hand, Kessel (1965, p. 25) and Cagan (1969, p. 93) give evidence intended to show that the liquidity premium increases with the level of rates. These tests have been called into question by Telser (1967) and McCulloch (1973, pp. 36–39).

Any direct comparison between our liquidity premium estimator and the level of rates is open to two objections. First, in deriving it, we assumed the underlying premium was constant. This can bias any comparison to the extent that the unforeseen change in rates has been correlated with the level of rates. Second, the premium estimator and the level are both calculated with measurement error from the same data. This can lend some inconsistency to any regression coefficients.

Although it should be interpreted with caution, the comparison is still worth making. Figure 1 and similar charts (not shown here) for $m =$ 3 months and $m = 16$ years show no obvious relation between our liquidity premium estimator and the level of interest rates. If there had been a pronounced relation, it would have been picked up by our maximum-likelihood test for nonconstancy of the premium since the Accord, in view of the fairly steady rise in interest rates over this period. Furthermore, direct regressions of the premium estimator on the level show no significant correlation either way for the post-Accord period.[15]

In view of the inconclusive nature of our own findings and the lack of consensus in the literature, we will assume that the post-Accord liquidity premium has been approximately independent of the level of interest rates.

III. The Post-Accord Liquidity Premium

Using the information we gathered in Section II about the behavior of the liquidity premium and the forecasting errors, we are ready to estimate the liquidity premium for a variety of maturities. We first estimate it without imposing any particular functional form on it. We find no evidence to contradict monotonicity or boundedness, so we then estimate it under the assumption that the premium monotonically approaches an asymptote. Since we are interested primarily in how free-market forces (relatively speaking) shape the term structure, and since the pre-Accord liquidity premium has been shown to be significantly different from that since the Accord, we restrict ourselves in this section to our 90 alternate post-Accord observations.

---

[15] See McCulloch (1973, p. 35). If period 1 is included in the regressions, the level has a positive and significant slope for $m = 3$ months. However, it is questionable whether this is the direct effect of the lower pre-Accord rates themselves, or if the low rates are only acting as a proxy for the higher level of intervention in the securities market prior to the Accord.

TABLE 5

FREE-FORM ESTIMATES OF LIQUIDITY PREMIUM
(POST-ACCORD EVEN OBSERVATIONS)

m	n	$\hat{\pi}(m)$	$\hat{\bar{\pi}}(m)$
0	0	0.0 (0.0)	0.0 (0.0)
1 month	90	0.19 (0.03)	0.09 (0.02)
2 months	90	0.32 (0.05)	0.17 (0.03)
3 months	90	0.36 (0.06)	0.23 (0.03)
6 months	90	0.37 (0.11)	0.31 (0.06)
9 months	90	0.32 (0.16)	0.32 (0.08)
1 year	90	0.33 (0.21)	0.32 (0.11)
2 years	90	0.34 (0.40)	0.33 (0.20)
3 years	90	0.37 (0.56)	0.33 (0.29)
5 years	90	0.47 (0.82)	0.39 (0.44)
10 years	90	−0.68 (1.40)	0.15 (0.75)
20 years	87	0.59 (2.10)	0.23 (1.24)
30 years*	67	−5.24 (3.81)	−2.47 (2.02)

NOTE.—Standard errors in parentheses.
* Periods 2 and 3 treated as if they had the same variance, due to shrinking sample size.

The Liquidity Premium and the Average Liquidity Premium

Table 5 gives our free-form estimates of the liquidity premium $\pi(m)$ and of the average liquidity premium,

$$\bar{\pi}(m) = \int_0^m \pi(x)\, dx. \tag{8}$$

The function $\pi(m)$ gives us the "typical" shape of the marginal forward interest rate curve $\rho(t, s)$, for if we expected to have the same marginal forward rates (as a function of maturity $m = s - t$) forever, we would find

$$\rho(t, s) = \rho(t, t) + \pi(s - t). \tag{9}$$

In this case, we would also expect to have the same yield curve forever and would have

$$\eta(t, m) = \eta(t, 0) + \bar{\pi}(m). \tag{10}$$

Thus, the average liquidity premium shown in table 5 gives us the "typical" shape of the yield curve. This average liquidity premium was estimated by averaging together the values

$$\bar{\pi}_j(m) = \frac{1}{m}\int_0^m \pi_j(x)\, dx, \tag{11}$$

again allowing for heteroskedasticity.[16]

The Mean Liquidity Premium

One of the most important applications of the liquidity premium is to evaluate the market's expectation of future interest rates spanning a positive interval in the future, on the basis of current forward rates. We define the mean forward rate observed at time t corresponding to a loan to begin at time $t + m_1$ and to be repaid at time $t + m_1 + m_2$ by

$$r(t, t + m_1, t + m_1 + m_2) = \frac{1}{m_2}\int_{m_1}^{m_1+m_2} \rho(t, t + m)\, dm. \tag{12}$$

In order to calculate the market's expectation of the future m_2-year yield to maturity $\eta(t + m_1, m_2)$ from this forward rate, we need to know the mean liquidity premium:

$$p(m_1, m_2) = r(t, t + m_1, t + m_1 + m_2) - E_t\eta(t + m_1, m_2). \tag{13}$$

It can be shown that

$$p(m_1, m_2) = [(m_1 + m_2)\bar{\pi}(m_1 + m_2) - m_1\bar{\pi}(m_1) - m_2\bar{\pi}(m_2)]/m_2. \tag{14}$$

Consequently,

$$\begin{aligned} p_j(m_1, m_2) = [(m_1 + m_2)\bar{\pi}_j(m_1 + m_2) - m_1\bar{\pi}_j(m_1) \\ - m_2\bar{\pi}_j(m_2)]/m_2 \end{aligned} \tag{15}$$

is an unbiased observation on $p(m_1, m_2)$. These observations were averaged together heteroskedastically to obtain the estimates of $p(m_1, m_2)$ shown in table 6.

It happens that when m_2 equals the observation interval Δt (1 month in our case), our estimator of the mean liquidity premium is identically equal to the difference between the 1-month holding period yield on a security of maturity $m + \Delta t$ and the certain yield on a security of Δt

---

[16] The integral in eq. (11) was evaluated by summing trapezoids with width $\Delta t =$ 1/12 year.

TABLE 6

FREE-FORM ESTIMATES OF MEAN LIQUIDITY PREMIUM $p(m_1, m_2)$

(Post-Accord Even Observations--Standard Errors in Parentheses)

m_2 m_1	0	1 mo.	2 mos.	3 mos.	6 mos.	9 mos.	1 year	2 years	3 years	5 years	10 yrs.	20 yrs.	30 yrs.
0	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
1 month	0.19 (0.03)	0.16 (0.02)	0.12 (0.02)	0.10 (0.02)	0.04 (0.02)	0.03 (0.02)	0.02 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.01)	-0.01 (0.01)	-0.00 (0.01)	-0.01 (0.01)
2 months	0.32 (0.05)	0.25 (0.04)	0.19 (0.03)	0.15 (0.04)	0.06 (0.04)	0.03 (0.04)	0.03 (0.03)	0.01 (0.03)	0.01 (0.03)	0.01 (0.03)	-0.01 (0.02)	-0.00 (0.02)	-0.03 (0.02)
3 months	0.36 (0.06)	0.29 (0.05)	0.22 (0.05)	0.16 (0.05)	0.06 (0.05)	0.04 (0.05)	0.03 (0.05)	0.02 (0.05)	0.01 (0.05)	0.01 (0.04)	-0.02 (0.03)	-0.00 (0.03)	-0.05 (0.03)
6 months	0.37 (0.11)	0.27 (0.11)	0.18 (0.11)	0.12 (0.11)	0.04 (0.11)	0.04 (0.11)	0.04 (0.10)	0.01 (0.10)	0.02 (0.09)	0.01 (0.08)	-0.05 (0.07)	-0.01 (0.06)	-0.10 (0.06)
9 months	0.32 (0.16)	0.23 (0.16)	0.15 (0.16)	0.11 (0.16)	0.05 (0.16)	0.05 (0.16)	0.05 (0.15)	0.01 (0.15)	0.02 (0.14)	0.01 (0.12)	-0.08 (0.10)	-0.01 (0.08)	-0.15 (0.10)
1 year	0.33 (0.21)	0.25 (0.21)	0.18 (0.21)	0.13 (0.21)	0.08 (0.21)	0.06 (0.20)	0.05 (0.20)	0.02 (0.20)	0.03 (0.18)	0.00 (0.16)	-0.11 (0.14)	-0.01 (0.11)	-0.20 (0.13)
2 years	0.34 (0.40)	0.24 (0.39)	0.17 (0.39)	0.12 (0.39)	0.05 (0.40)	0.04 (0.39)	0.03 (0.39)	0.03 (0.38)	0.05 (0.35)	-0.07 (0.32)	-0.24 (0.27)	0.04 (0.23)	-0.42 (0.25)
3 years	0.37 (0.56)	0.28 (0.56)	0.21 (0.56)	0.16 (0.56)	0.10 (0.55)	0.09 (0.55)	0.09 (0.55)	0.08 (0.53)	0.05 (0.51)	-0.20 (0.48)	-0.38 (0.40)	0.09 (0.34)	-0.53 (0.38)
5 years	0.47 (0.82)	0.37 (0.82)	0.29 (0.81)	0.23 (0.81)	0.12 (0.82)	0.07 (0.81)	0.02 (0.81)	-0.17 (0.80)	-0.34 (0.80)	-0.65 (0.79)	-0.66 (0.64)	-0.83 (0.62)	-1.28 (0.68)
10 years	-0.68 (1.40)	-0.77 (1.40)	-0.86 (1.40)	-0.92 (1.40)	-1.02 (1.40)	-1.07 (1.39)	-1.09 (1.39)	-1.21 (1.36)	-1.26 (1.34)	-1.33 (1.27)	-0.48 (1.16)	-1.91 (1.31)	--
20 years	0.59 (2.10)	-0.14 (2.24)	-0.21 (2.24)	-0.25 (2.25)	-0.30 (2.25)	-0.28 (2.24)	-0.24 (2.23)	0.36 (2.31)	0.60 (2.27)	-3.31 (2.49)	-3.82 (2.62)	--	--
30 years	-5.24 (3.81)	-5.38 (3.81)	-5.52 (3.81)	-5.62 (3.81)	-5.84 (3.81)	-5.95 (3.80)	-6.03 (3.79)	-6.29 (3.75)	-5.30 (3.81)	-7.67 (4.08)	--	--	--

(McCulloch 1973, pp. 43–45). Thus, our estimator is the same as Cagan's difference in holding period yields, except that we use an observation interval of 1 month, while he used 1 week. By developing it as we have, however, we are able to estimate the typical shape of the yield and forward curves exactly.[17]

Transactions Costs

If there were no transactions costs, it would follow from table 6 that for many pairs of maturities, it is worth the while of a lender who desires a short holding period to buy a longer maturity and to sell it before maturity, rather than simply to buy the shorter maturity in the first place. However, he cannot really both buy and sell at the bid-asked mean prices as implicitly assumed so far.

We cannot quantify all the components of the cost to an investor or borrower of going in and out of securities. However, the quoted bid-asked spread gives us a fair estimate of the external component of this cost, at least for large institutional investors. Subject, then, to the condition that he must buy high at the asked price and sell low at the bid price, we may investigate whether the short-term lender does better or worse to buy a longer maturity initially and sell it before maturity than simply to buy a short security.

We found that a lender who wished to lend for only 1, 2, or 3 months would have done significantly better to have bought a 2-, 4-, or 6-month security, respectively, even though it would mean having to go into the market twice instead of only once (McCulloch 1973, pp. 46–50). When we compared these differentials with the comparable values from table 6, we found that the bid-asked spread eats up only about one-third of the liquidity premium for these three pairs of maturities. However, the essentially short-term nature of the liquidity premium is evidenced by the fact that a lender who wished to lend for 4 months would not have done significantly better to have bought an 8-month security.

Miller and Orr (1967, pp. 133–51) have estimated the cost to one large nonfinancial corporation of going into and out of a given security as about $20–$50. Half this internal transaction cost just equals the premium (after bid-asked spread cost) on a 6-month bill held for 3 months, for investments of $36,000–$91,000. Therefore, a transaction would have to be in the hundreds to thousands of dollars before it definitely pays to try to exploit the liquidity premium.

---

[17] Culbertson (1957, p. 506) also works in terms of the difference in holding-period yields. Roll (1970, pp. 98–99) estimates the liquidity premium by a sum of differences of forward rates that is also equal to the difference in holding-period yields. In McCulloch (1973, p. 45), we show that interpolating from our table 6 for $m_1 = 6$ months and $m_2 = 1$ week gives an estimate not significantly different from Roll's.

Exponential Form Estimates of the Liquidity Premium

In tables 5 and 6 we made no assumptions about the form of the liquidity premium as a function of m. Since we put relatively little in, we got relatively little out in return. Thus, for longer maturities, where the variance of the forecasting errors is high, our estimators have very high standard errors. To return to the example we cited in Section I, the forward rate for a 10-year loan 5 years in the future from June 30, 1953 was 3.16 percent per annum. From table 6 ,we see that the premium on this rate would be -0.66 (± 0.64) percent per annum, so that a 95 percent confidence interval for the market expectation of the future 10-year spot rate extends from 2.54 percent per annum to 5.10 percent per annum (2.0 standard errors in each direction from $3.16 - [-0.66]$). The subsequently observed spot rate was 2.94 percent per annum, which lies within this interval, so we may not say for certain whether there was an unanticipated rise or fall, or no unanticipated change at all.

In order to pin the liquidity premium down with greater precision, we must make stronger assumptions. Kessel (1965) has argued that the liquidity premium should be such as to give the "typical" shape of the yield curve a monotonically increasing shape that approaches a horizontal asymptote.[18] If so, this information would eliminate much of our present uncertainty as to the behavior of $\pi(m)$ for m greater than 1 year or so.

Our free-form estimate of $\pi(m)$ rises in the first few months and then levels off. Although it falls off after 5 years, its final value (-5.24 ± 3.81 percent per year) is not significantly lower than its highest value (0.59 ± 2.10 percent per year), so its downward slope for long maturities is not necessarily significant.[19] The low values for long maturities may simply reflect the rise in interest rates over the period studied, which was probably largely unanticipated.

Since the observed liquidity premium seems not to differ significantly from the monotonically increasing-to-an-asymptote shape, we will try fitting it under that assumption. A simple two-parameter function that has these properties is

$$\pi(m) = b(1 - e^{-am}). \tag{16}$$

The parameters a and b were fit to our data by the maximum-likelihood technique as follows: free-form estimates of $\pi(m)$ for the six maturities 1 month, 3 months, 6 months, 1 year, 5 years, and 15 years were obtained as in table 5. Then intermaturity covariances were estimated about these means, producing a different 6 × 6 covariance matrix for each of the three post-Accord periods. The likelihood function was computed and maximized with respect to a and b based on the assumption that each set of six $\pi_j(m)$ constituted a multinormal drawing about the means given in (16). Given a, the maximizing value of b can be written in closed form. Different values of a were tried, until upper and lower bounds were found within .0005 of one another. The final estimator of a was placed halfway between these bounds, and the estimator of b calculated for this $\hat{a}$. The covariance matrix of the estimators of a and b was estimated by evaluating at $\hat{a}$ and $\hat{b}$ the second partial derivatives with respect to a and b of the logarithm of the likelihood function and inverting the negative of the matrix of these derivatives (Goldberger 1964, p. 131). The estimates thus obtained and their standard errors and covariance are

---

[18] Hicks (1946, p. 147) argues in favor of a monotonically increasing liquidity premium due to risk aversion in the face of interest-rate uncertainty. However, this is a non sequitur (Long 1972 and McCulloch 1973, pp. 4–10). Nevertheless, we give two alternative justifications of Kessel's monotonic shape, based on considerations involving the uncertain timing of receipts and expenditures and on the nature of a fractional reserve banking system (McCulloch 1973, pp. 10–15).

[19] A more powerful test of monotonicity is performed below.

$$\begin{aligned} \hat{a} &= \underset{(1.068)}{6.059}(\text{yr})^{-1}, \\ \hat{b} &= \underset{(0.0738)}{0.4335} \text{ percent per year}, \\ \hat{\text{cov}}(\hat{a}, \hat{b}) &= 0.06262. \end{aligned} \tag{17}$$

Given our formula for $\pi(m)$, we may compute $\bar{\pi}(m)$ and $p(m_1, m_2)$ by:

$$\bar{\pi}(m) = b[1 - (1 - e^{-am})/(am)], \tag{18}$$

$$p(m_1, m_2) = b(1 - e^{-am_1})(1 - e^{-am_2})/(am_2). \tag{19}$$

Standard errors for these nonlinear functions of a and b were approximated using the formula for asymptotic variances (Goldberger 1964, pp. 122–25). The results are shown in tables 7 and 8. Figures 2 and 3 show typical forward curves and yield curves incorporating the estimated values of $\pi(m)$ and $\bar{\pi}(m)$. Figure 2 uses the free-form estimates, while figure 3 uses the exponential-form estimates.

Table 6 can be used to answer a number of questions. The first column of figures tells us how much more return we would make, on average, by turning over daily in securities of maturity m than we would by turning over in "call money." (These comparisons ignore transactions costs, which of course would make it prohibitively expensive to buy a 30-year bond, hold it for a day, and then replace it by a new bond with a full 30 years to run. Our "call money" rate is the limit as maturity goes to zero of the bid-asked mean rate on Treasury bills.) The second column tells us how much more we would make by holding securities of maturity m to maturity than we would make by turning over in call money. By subtraction we also obtain the difference in yield to maturity we would obtain over a given period by holding securities of different maturities.

TABLE 7

EXPONENTIAL-FORM ESTIMATE OF LIQUIDITY PREMIUM (POST-ACCORD EVEN OBSERVATIONS)

m	$\hat{\pi}(m)$	$\hat{\bar{\pi}}(m)$	$\hat{\pi}(\infty) - \hat{\pi}(m)$	$\hat{\bar{\pi}}(\infty) - \hat{\bar{\pi}}(m)$
0	0.000	0.000	0.433	0.433
	(0.000)	(0.000)	(0.074)	(0.074)
1 month	0.172	0.093	0.262	0.340
	(0.018)	(0.010)	(0.065)	(0.069)
2 months	0.276	0.161	0.158	0.273
	(0.030)	(0.017)	(0.052)	(0.064)
3 months	0.338	0.210	0.095	0.223
	(0.040)	(0.023)	(0.040)	(0.058)
6 months	0.413	0.297	0.021	0.136
	(0.062)	(0.037)	(0.014)	(0.041)
9 months	0.429	0.339	0.005	0.094
	(0.070)	(0.046)	(0.004)	(0.030)
1 year	0.432	0.362	0.001	0.071
	(0.073)	(0.052)	(0.001)	(0.023)
2 years	0.433	0.398	0.000	0.036
	(0.074)	(0.063)	(0.000)	(0.012)
3 years	0.433	0.410	0.000	0.024
	(0.074)	(0.066)	(0.000)	(0.008)
5 years	0.433	0.419	0.000	0.014
	(0.074)	(0.069)	(0.000)	(0.005)
10 years	0.433	0.426	0.000	0.007
	(0.074)	(0.072)	(0.000)	(0.002)
20 years	0.433	0.430	0.000	0.004
	(0.074)	(0.073)	(0.000)	(0.001)
30 years	0.433	0.431	0.000	0.002
	(0.074)	(0.073)	(0.000)	(0.001)

NOTE.—Standard errors in parentheses.

For example, by holding 3-year notes to maturity, we would expect to get 0.410 − 0.210 = 0.200 percent per year more than we would by holding a sequence of 3-month bills to maturity. The standard error of this figure depends on a covariance not shown, but it must lie between |0.066 + 0.023| = 0.089 and |0.066 − 0.023| = 0.043.

The third column of figures in table 6 tells us how much more we would make by turning over daily in long maturities ($m \rightarrow \infty$) than by turning over daily in maturity m. Note that on this score there is no perceptible difference between 2-year bonds and perpetuities. The last column tells us how much more we get if we hold longs to maturity than if we hold securities of maturity m to maturity. Comparable free-form estimates do not appear in table 5 because without our exponential-form assumption (or something comparable), we have no information about what happens at infinity.

Our exponential-form estimates are never significantly different from our free-form estimates. Indeed, it is only rarely that they differ by more than 1 free-form standard error. However, our assumption of the exponential form greatly reduces the standard errors. The premium on

TABLE 8

EXPONENTIAL-FORM ESTIMATES OF MEAN LIQUIDITY PREMIUM $p(m_1, m_2)$

(Post-Accord Even Observations--Standard Errors in Parentheses)

m_2 m_1	0	1 mo.	2 mos.	3 mos.	6 mos.	9 mos.	1 year	2 years	3 years	5 years	10 yrs.	20 yrs.	30 yrs.
0	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
1 month	0.17 (0.02)	0.13 (0.01)	0.11 (0.01)	0.09 (0.01)	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
2 months	0.28 (0.03)	0.22 (0.03)	0.17 (0.03)	0.14 (0.02)	0.09 (0.02)	0.06 (0.01)	0.05 (0.01)	0.02 (0.01)	0.02 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
3 months	0.34 (0.04)	0.27 (0.04)	0.21 (0.04)	0.17 (0.03)	0.11 (0.02)	0.07 (0.02)	0.06 (0.01)	0.03 (0.01)	0.02 (0.00)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)
6 months	0.41 (0.06)	0.32 (0.06)	0.26 (0.05)	0.21 (0.05)	0.13 (0.04)	0.09 (0.03)	0.07 (0.02)	0.03 (0.01)	0.02 (0.01)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)
9 months	0.32 (0.07)	0.34 (0.07)	0.27 (0.06)	0.22 (0.06)	0.13 (0.04)	0.09 (0.03)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)
1 year	0.43 (0.07)	0.34 (0.07)	0.27 (0.06)	0.22 (0.06)	0.14 (0.04)	0.09 (0.03)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)
2 years	0.43 (0.07)	0.34 (0.07)	0.27 (0.06)	0.22 (0.06)	0.14 (0.04)	0.09 (0.03)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)
3 years	0.43 (0.07)	0.34 (0.07)	0.27 (0.06)	0.22 (0.06)	0.14 (0.04)	0.09 (0.03)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)
5 years	0.43 (0.07)	0.34 (0.07)	0.27 (0.06)	0.22 (0.06)	0.14 (0.04)	0.09 (0.03)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)
10 years	0.43 (0.07)	0.34 (0.07)	0.27 (0.06)	0.22 (0.06)	0.14 (0.04)	0.09 (0.03)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)
20 years	0.43 (0.07)	0.34 (0.07)	0.27 (0.06)	0.22 (0.06)	0.14 (0.04)	0.09 (0.03)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)
30 years	0.43 (0.07)	0.34 (0.07)	0.27 (0.06)	0.22 (0.06)	0.14 (0.04)	0.09 (0.03)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)

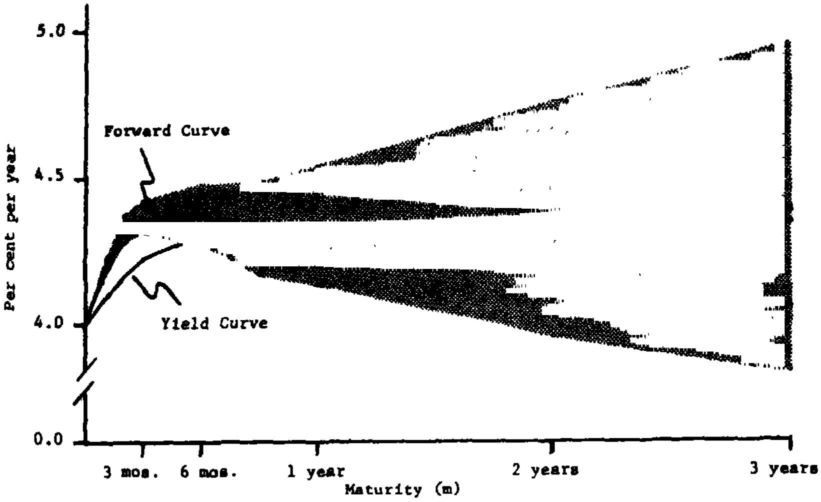

FIG. 2.—Typical shape of forward curve and yield curve under free-form assumption. Short rate assumed to be 4 percent per year. Forward curve shown with band extending ±1 standard error.

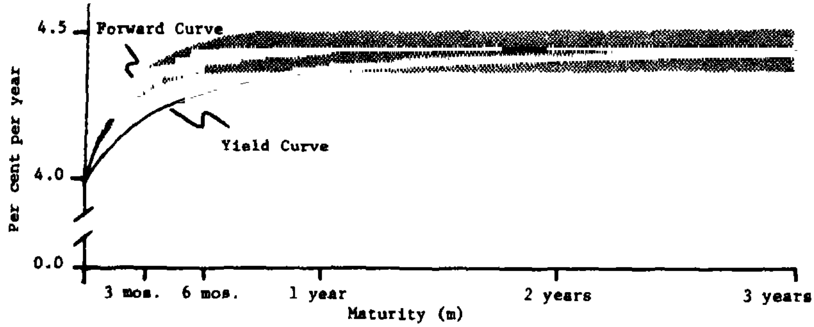

FIG. 3.—Typical shape of forward curve and yield curve under exponential form assumption. Short rate assumed to be 4 percent per year. Forward curve shown with band extending ±1 standard error.

a 10-year rate 5 years forward is now 0.01 (± less than 0.005). Our estimate of the June 30, 1953 forecast of the June 30, 1958 10-year spot rate becomes 3.16 − 0.01 = 3.15 percent per annum. The measurement errors of the forward rate and the premium combine to give a standard error of about 0.01, so our 95 percent confidence interval now extends from 3.13 to 3.17 percent per year. The subsequent spot rate was 2.94 (±0.02) percent, so we may definitely say that the fall was unanticipated—provided, of course, we are willing to accept the assumptions of monotonicity and boundedness that led us to the exponential form.

The striking aspect of table 8 is the very low values for the mean liquidity premium when m_2 (the duration of the forward loan) is larger than a year or so. Previous investigations have given the impression that the liquidity premium is on the order of 0.5 percent per year. Our results are not inconsistent with these levels, provided m_2 is very small, as it has been in other studies. When m_2 becomes large, however, the premium in the subsequent spot rate is almost as large as that in the forward rate. The two nearly cancel out, leaving only a small residual. If we make no assumptions about the form of the premium, as in table 6, we get a very large confidence interval that gives us little information, due to the large amount of "noise" at these maturities. But then we impose monotonicity and boundedness, the additional information closes the confidence interval about a value very near zero.

Under our exponential assumption, $p(m_1, m_2)$ as defined in (19) obeys the following inequality:

$$p(m_1, m_2) < (b/a)/m_2. \tag{20}$$

This inequality gives us a convenient upper bound for the liquidity premium. The value of b/a is 0.072 percent (standard error = 0.023), so we may state at the 95 percent confidence level that the mean premium is less than $0.101/m_2$ percent per year, when m_2 is measured in years.[20] If m_2 is a small fraction of a year, the liquidity premium can be substantial. But when m_2 is larger than 3 years or so, the premium is less than the precision with which forward rates can be measured.[21]

We are able to test the exponential-form estimates against the free-form estimates by means of the likelihood ratio test. With the free form we are estimating six independent means. With the exponential form, we are constraining these six values to conform to a two-parameter function, so twice the logarithm of the likelihood ratio should be compared with the χ_4^2 distribution.[22] Since its value was 4.71, which is not significant at even the 70 percent level, we may not reject the exponential form. Incidentally, this test provides a test for monotonicity. Since we may not reject this specific monotonic form, it follows a fortiori that we may not reject monotonicity in general.

---

[20] For a one-tailed test, the boundary of the 95 percent confidence region is approximately 1.70 standard errors above the mean.

[21] The logit form $\pi(m) = bm/(a + m)$ was tried in addition to the exponential form. Even though it increases monotonically to an asymptote, it does not imply an upper bound such as that given in eq. (20). However, the exponential form gave a slightly higher likelihood and therefore was preferred.

[22] The relevant free-form estimates for this test are ones which maximize the joint likelihood, taking into account the intermaturity covariances. These estimates are slightly different from those given in table 5, which take each maturity by itself.

Comparison with Survey Study

A survey of market expectations conducted by Kane and Malkiel[23] on April 1, 1965, affords us an opportunity to check our liquidity premium estimates against estimates that do not depend on our assumption of unbiased forecasting. Their sample consisted of 200 banks, life insurance companies, and nonfinancial corporations. Respondents varying in number from 77 to 90 ventured opinions as to the 90-day bill rate and 10-year bond rate on the various future dates listed in table 9. The standard errors of the mean market forecasts are remarkably small. For the 90-day-bill forecasts, they are actually considerably smaller than our measurement errors for the corresponding forward rates.

The survey premia are never significantly different from our free-form estimates from table 6, thanks to the high errors on the latter. And the survey premia are not significantly different from our exponential-form estimates for loans of 90 days' duration. In large measure this is due to the ambiguity of the forward rates. However, we run into trouble when we try to compare the survey premia on 10-year bonds to our table 8 estimates. For the bond 1 year in the future, the survey premium exceeds ours by four basis points or 1 standard error, and for 2 years, by 21 basis points, or 4 standard errors. This discrepancy casts some doubt on the validity of either our exponential form or the survey responses. It may be that the exponential form approaches its asymptote too quickly. In defense of our results, however, it should be noted that the survey results (taken together with the forward rates we have computed) strongly suggest that p(2 years, 10 years) is about four times p(1 year, 10 years), contradicting the widely accepted hypothesis that the premium increases with distance into the future at a decreasing rate.[24] The comparison of survey estimates of expectations with unbiased forecasting estimates like ours appears to be a fruitful field for future research.

IV. Conclusion

Without imposing any particular form on the liquidity premium, we were able to demonstrate the following: there is a liquidity premium, significantly greater than zero. This premium has been large enough since the Accord to imply that for some maturities, borrowers or lenders who desire one borrowing or lending period would do better in a different maturity, in spite of the extra costs incurred. The premium has been larger since the Accord than it was during the 4 years before the Accord.

---

[23] (1967). I am grateful to Malkiel for a letter explaining some of the figures given in that paper.

[24] The problem cannot be that the questionnaires were not filled out promptly on April 1, but only after a delay of variable length, for the forward rates in question were actually higher on April 30 and May 31 than they had been at the close of March.

TABLE 9

COMPARISON OF PREMIA FROM TABLES 6 AND 8 WITH THOSE DERIVED FROM KANE-MALKIEL SURVEY

Future Date	Security	Forward Rate*	Survey Expected Rate†	Survey Premium	Exponential-Form Premium	Free-Form Premium
July 1, 1965	90-day bill	4.07	3.99	.06	.17	.16
		(.13)	(.01)	(.13)	(.03)	(.05)
October 1, 1965	90-day bill	4.04	4.01	.03	.21	.12
		(.18)	(.01)	(.18)	(.05)	(.11)
January 1, 1966	90-day bill	3.84	3.96	−.12	.22	.11
		(.18)	(.02)	(.18)	(.06)	(.16)
April 1, 1966	90-day bill	3.81	3.86	−.05	.22	.13
		(.19)	(.02)	(.19)	(.06)	(.21)
April 1, 1967	90-day bill	3.93	3.66	.27	.22	.12
		(.10)	(.04)	(.11)	(.06)	(.39)
April 1, 1966	10-year bond	4.17	4.12	.05	.01	−.11
		(.03)	(.02)	(.04)	(.00)	(.14)
April 1, 1967	10-year bond	4.20	3.98	.22	.01	−.24
		(.04)	(.03)	(.05)	(.00)	(.27)

NOTE.—Standard errors in parentheses.

* Rates computed from U.S. Government security prices as of March 31, 1965, by method of McCulloch (1971). Logarithmic rates converted to banker's discount basis for comparison to bills.

† Mean of "most likely" rate for all respondents. Standard error equals standard deviation given by Kane and Malkiel (1967), divided by the square root of the number of respondents.

Finally, the variance of forecasting errors has not been constant over the period since the Accord.

We were unable to demonstrate any variation in the liquidity premium itself since the Accord, either as a function of time or as a function of the level of interest rates. We were also unable to detect nonnormality in the distribution of the forecasting errors (provided we allow for heteroskedasticity), or nonmonotonicity in the behavior of the premium as a function of maturity.

However, if we do not impose a particular form, our estimates of the mean term premium on long-term forward rates are very inaccurate. When we postulate the monotonic and bounded exponential form of (16) we obtain the relatively precise estimates contained in tables 7 and 8. In particular, we may then say that the forward rate corresponding to a future loan of duration m years is greater than the expected future spot rate, but by no more than $0.101/m$ percent per year.

References

Abbott, Charles C. "A Note on the Government Bond Market, 1919–1930." *Rev. Econ. and Statis.* 17 (January 1935): 7–12.

Bierwag, G. O., and Grove, M. A. "A Model of the Term Structure of Interest Rates." *Rev. Econ. and Statis.* 40 (1967): 50–62.

Bryan, William R., and Carleton, Willard T. "A Method of Estimating the Term Structure of Interest Rates." Unpublished paper given at meetings of the American Finance Association, Toronto, December 1972.

Buse, Adolf. "Expectations, Prices, Coupons, and Yields." *J. Finance* 25 (September 1970): 809–18.

Cagan, Phillip. "A Study of Liquidity Premiums on Federal and Municipal Government Securities." In *Essays on Interest Rates*, vol. 1, edited by J. M. Guttentag and P. Cagan. New York: Nat. Bur. Econ. Res., 1969.

Culbertson, John M. "The Term Structure of Interest Rates." *Q.J.E.* 71 (November 1957): 485–517.

David, H. A.; Hartley, H. O.; and Pearson, E. S. "The Distribution of the Ratio, in a Single Normal Sample, of Range to Standard Deviation." *Biometrika* 41 (1954): 482–93.

Diller, Stanley. "Expectations in the Term Structure of Interest Rates." *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance.* Edited by J. Mincer. New York: Nat. Bur. Econ. Res., 1969.

Fama, Eugene F., and Roll, Richard. "Some Properties of Symmetric Stable Distributions." *J. American Statis. Assoc.* 63 (September 1968): 817–36.

———. "Parameter Estimates for Symmetric Stable Distributions." *J. American Statis. Assoc.* 66 (June 1971): 331–38.

Goldberger, Arthur S. *Econometric Theory.* New York: Wiley, 1964.

Hart, B. I. "Tabulation of the Probabilities for the Ratio of the Mean Square Successive Difference to the Variance." *Ann. Math. Statis.* 13 (1942): 207–14.

Hicks, John R. *Value and Capital.* 2d ed. Oxford: Clarendon Press, 1946.

Kane, Edward J., and Malkiel, Burton G. "The Term Structure of Interest Rates: An Analysis of a Survey of Interest Rate Expectations." *Rev. Econ. and Statis.* 69 (August 1967): 343–55.

Kessel, Reuben A. *The Cyclical Behavior of the Term Structure of Interest Rates*. New York: Nat. Bur. Econ. Res., 1965. Reprinted in *Essays on Interest Rates*. Vol. 2. Edited by J. M. Guttentag. New York: Nat. Bur. Econ. Res., 1971.

Long, John B., Jr. "Consumption-Investment Decisions and Equilibrium in the Securities Market." In *Studies in the Theory of Capital Markets*, edited by M. C. Jensen. New York: Praeger, 1972.

McCulloch, J. Huston. "Measuring the Term Structure of Interest Rates." *J. Bus.* 44 (January 1971): 19–31.

———. "An Estimate of the Liquidity Premium." Ph.D. dissertation, Univ. Chicago, 1973.

Meiselman, David. *The Term Structure of Interest Rates*. Englewood Cliffs, N.J.: Prentice-Hall, 1962.

Miller, Merton H., and Orr, Daniel. "An Application of Control-Limit Models to the Management of Corporate Cash Balances." *Financial Research and Management Decisions*. Edited by A. A. Robichek. New York: Wiley, 1967.

Modigliani, Franco, and Shiller, Robert J. "Inflation, Rational Expectations, and the Term Structure of Interest Rates." *Economica*, n.s. 40 (February 1973): 12–43.

Modigliani, Franco, and Sutch R. "Innovations in Interest Rate Policy." *A.E.R.* 56 (May 1966): 178–97.

Nelson, Charles R. *The Term Structure of Interest Rates*. New York: Basic Books, 1972.

Praetz, Peter D. "The Distribution of Share Price Changes." *J. Bus.* 45 (January 1972): 49–55.

Press, S. James. "A Compound Events Model for Security Prices." *J. Bus.* 40 (July 1967): 317–35.

Rice, John R. *The Approximation of Functions*. Vol. 2. Reading, Mass.: Addison-Wesley, 1969.

Roll, Richard. *The Behavior of Interest Rates: The Application of the Efficient Market Model to U.S. Treasury Bills*. New York: Basic Books, 1970.

Samuelson, Paul A. "Proof That Properly Anticipated Prices Fluctuate Randomly." *Indus. Management Rev.* 6 (Spring 1965): 41–49.

Stiglitz, Joseph E. "A Consumption-oriented Theory of the Demand for Financial Assets and the Term Structure of Interest Rates." *Rev. Econ. Studies* 37 (July 1970): 321–51.

Telser, Lester G. "A Critique of Some Recent Empirical Research on the Explanation of the Term Structure of Interest Rates." *J.P.E.* 75, no. 4, pt. II (August 1967): 546–68.

Van Horne, James. "Interest Rate Risk and the Term Structure of Interest Rates." *J.P.E.* 73, no. 4 (August 1965): 344–51.

Wallace, Neil. "The Term Structure of Interest Rates and the Maturity Composition of the Federal Debt." Ph.D. dissertation, Univ. Chicago, 1964.

Weingartner, H. Martin. "The Generalized Rate of Return." *J. Financial and Quantitative Analysis* 1 (September 1966): 1–29.

Williams, John Burr. *The Theory of Investment Value*. Cambridge, Mass.: Harvard Univ. Press, 1938.

Trade in Technological Knowledge and the National Advantage

Carlos Alfredo Rodriguez

Columbia University

The optimal trade and licensing policies for a country which is the only owner of the technology to produce a good are developed in a general-equilibrium two-countries model. The analysis emphasizes the monopolistic position of the owner of the technology, and is essentially static in the sense that no process of generation of new technology is considered. The optimal behavior obtained turns out to entail (*a*) the full exploitation of any monopoly and/or monopsony power that the owner of the technology may possess with respect to the foreign market for his product and the foreign market for the factors which may be used in its production, and (*b*) full, competitive behavior with respect to the use of domestic resources and the sale of his product in the domestic market.

Most models of international trade focus on exchange of goods or factors of production. Technology is usually assumed to be the same for each country. However, the production of a traded good in a particular country may reflect not the differences in factor endowments which are central to the Heckscher-Ohlin model, but differences in technology or "know-how" across countries. Evidence supporting this possibility is the increasing importance of international patent agreements, which control the international transfer of new technology. This paper analyzes the optimum policy for a country which owns the technology for producing a unique product (Coca Cola?). Because I disregard the effects of pricing policies on the generation of new technology, the analysis is essentially static. A dynamic model, linking the rental for use of existing technology

This paper is a revised version of one chapter of my Ph.D. dissertation ("Aspects of International Factor Mobility," University of Chicago, 1973). In addition to the members of my dissertation committee, I would like to thank Jacob Frenkel, Harry G. Johnson, and an anonymous referee for helpful comments and suggestions.

[*Journal of Political Economy*, 1975, vol. 83, no. 1]

© 1975 by The University of Chicago. All rights reserved.

to the rate of generation of comparable or new technological knowledge, should be a matter for further investigation.

I assume that, once the technology is known, only ordinary capital and labor are required to produce the good. Transmission of technology between countries is assumed costless. Thus, it is possible for the country which owns the technology to operate a plant in a foreign country without any transfer of factors. This contrasts with the well-known models of Jones (1967) and Kemp (1966), in which foreign investment is identified with the transfer of physical capital. In the Kemp-Jones models, a tariff and a tax on capital movements are used to maximize welfare of the home country; here, national welfare is maximized by a tariff and a rent charged for the use of the technology by (or in) the foreign country.

Policies available to the home country include trade restrictions (tariffs), licensing of foreign producers with per-unit or lump-sum royalty payments, opening a foreign production facility, and exploiting any monopoly or monopsony power in foreign markets. Combinations of these policies are also possible, for example, licensing foreign producers but also imposing a tariff. In Section I, I present a geometric analysis of these choices in the simple model of two countries with identical constant costs. In this case an optimal policy for the home country is to license foreign producers, using a per-unit royalty. An alternative strategy of exporting to the foreign country and charging an optimal export tariff—which turns out to be the same as the optimal royalty rate—will yield the same welfare maximum as long as the home country does not become specialized in production. In Sections II and III, I analyze some of the alternatives algebraically, while allowing for increasing and different relative costs in the two countries.

Market structure affects both the welfare optimum for the home country and the means by which it is implemented. I consider only two cases, which turn out to be identical in the nature of the optimum but different in the method of implementation. In Sections I and II it is assumed that competition prevails at home and the technology is licensed to competitive foreign producers. In Section III, I consider the case in which the good is produced in both countries by a domestic firm which has monopoly and/or monopsony power.

I

I assume two countries: Country I (the home country, exclusive owner of the technology to produce soft drinks) and Country II (the foreign country). There are two goods, a soft drink (*D*) and wheat (*W*), produced under constant costs. The dashed line in figure 1 represents the potential production possibility frontier (PPF) of Country II when it has access to the technology for producing *D*. The assumption of identical constant

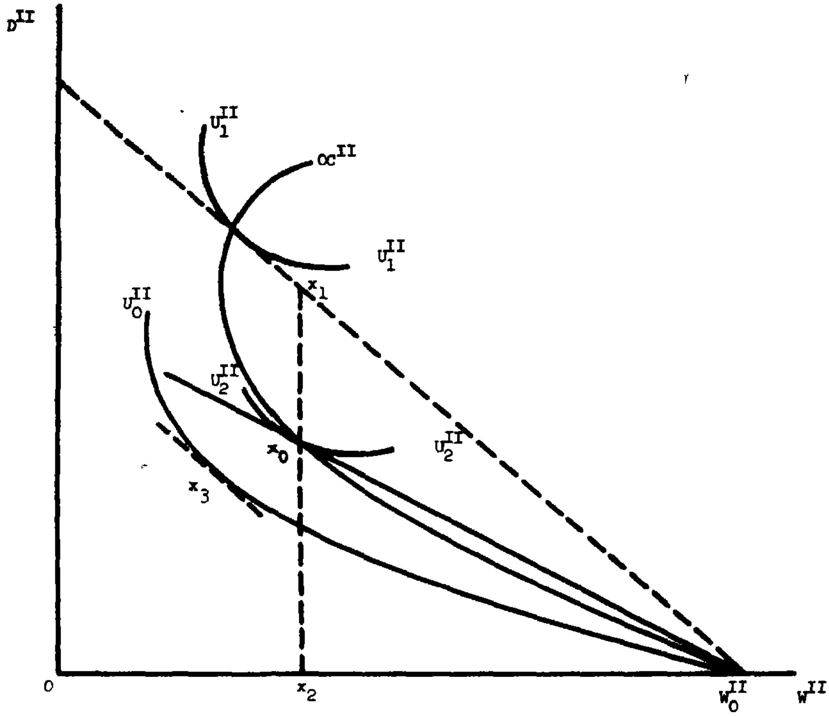

Fig. 1

costs in the two countries implies that the PPF is identical with that of Country I except for scale. In the absence of licensing or the opening of a foreign branch by Country I, Country II lacks the technology to produce D, so its PPF consists of the single point W_0^{II} on the wheat axis.

Without trade or licensing, Country II produces and consumes at W_0^{II} in figure 1 and enjoys a level of utility corresponding to the indifference curve U_0^{II}. Country I produces and consumes at $(D_0^{\text{I}}, W_0^{\text{I}})$ on the indifference curve U_0^{I} in figure 2. (In what follows I assume for simplicity that the indifference map of Country I is homothetic.) With free trade, Country I utility remains at the autarky level as long as Country I is incompletely specialized. Consumption of each good is unchanged, while production shifts to $(D_1^{\text{I}}, W_1^{\text{I}})$ in response to the trading demands of Country II. Country II gains as much under free trade as it would from free access to the technology for producing D. Its utility level is now U_1^{II}, the indifference curve tangent to the potential PPF. That only Country II gains from trade is a consequence of the assumptions of constant costs and nonspecialization in Country I.

We can now analyze the effects of different trade and licensing policies on the welfare of the home country. I assume initially that Country I

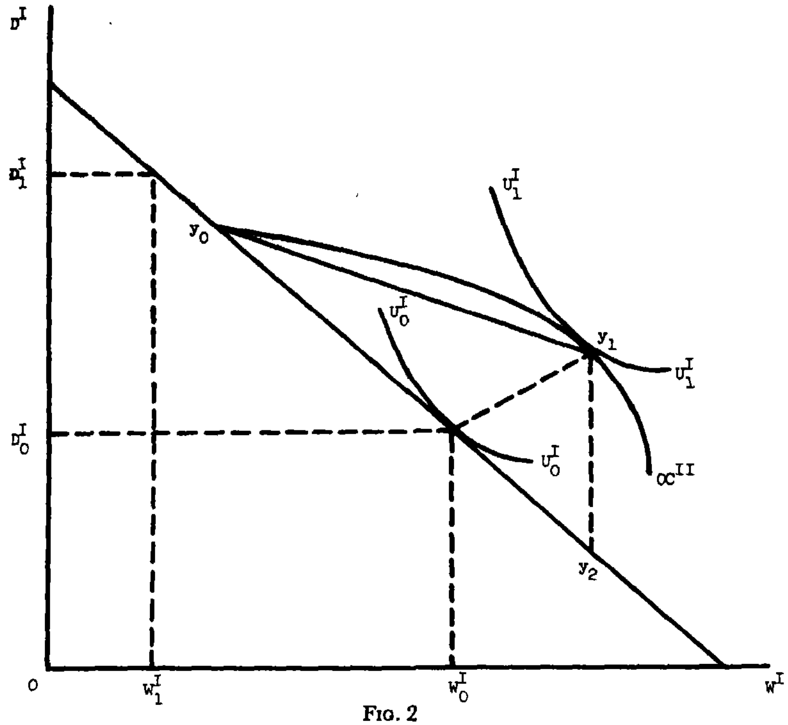

FIG. 2

does not specialize in production of D. Suppose that Country I imposes an optimal tariff t^* on its trade with Country II. This situation is also depicted in figure 2. The Country I consumption possibility frontier is the envelope derived by moving the origin of the foreign offer curve OC^{II} through each possible production point along the Country I PPF. (Each point on the offer curve OC^{II} is the tangency of a price line through the endowment point W_0^{II} in figure 1 and a Country II indifference curve). The Country I optimum consumption point y_1 is given by the tangency of a Country I indifference curve U_1^I with the consumption possibility frontier and, therefore, with the Country II offer curve, OC^{II}. The corresponding production point is y_0, and the terms of trade are given by the slope of the line y_0y_1, which is equal to the slope of $x_0W_0^{II}$ in figure 1.

The rate of tariff—in this case an export tariff on D paid on the basis of the gross price received by the exporter—is the ratio x_0x_1/x_1x_2 in figure 1. Country II utility is reduced to U_2^{II}, which, however, is still higher than under autarky.

If, instead, the home country licenses its technology to competitive foreign producers and an optimal royalty rate r^* is charged per unit of production, the outcome will be basically identical with that of trade with the optimal tariff. It is easy to see that the royalty rate which maximizes royalty payments is x_0x_1/x_1x_2 and that total royalties collected are measured by x_0x_1. Country I must impose an export tariff at the same rate as the royalty to prevent Country II consumers from purchasing *D* at the lower price in the home country. With the optimal royalty, Countries I and II consume at the same levels as under the optimal tariff. Country I now produces at y_2 and collects royalties given by the distance y_1y_2. Country II produces at x_1 and pays $x_0x_1 = y_1y_2$ in royalties to Country I. Given the export tariff in Country I, there is no incentive for trade.

A third possibility is that producers of *D* in Country I open a production facility abroad and behave as monopolists. (The constant-costs assumption rules out any monopsony power.) The outcome will again be basically identical; the price which maximizes profits is the same as that obtained under the tariff or royalty policies. In this case, the distance $y_1y_2 = x_0x_1$ represents profits from investments of Country I abroad. Under this alternative, Country I again needs an export tariff because of the higher foreign price of *D*.

We can thus conclude that with identical constant costs and incomplete specialization in the home country, trade with an optimal tariff, licensing with an optimal royalty per unit produced, or opening a branch abroad and charging a monopoly price are perfect substitutes from the point of view of both countries.

I now turn to the case of a lump-sum royalty payment per unit of time. As seen in figure 1, the slope of the PPF at the production point x_1 is different from the marginal rate of substitution in consumption at x_0. From this it is easy to infer that there is a lump-sum royalty payment which yields more revenue to Country I while keeping Country II at the same utility level. The maximum lump-sum royalty payment which can be extracted from Country II is found by locating the point at which the slope of the autarky indifference curve U_0^{II} is equal to that of the PPF; this occurs at x_3, and the maximum lump-sum royalty is the vertical distance between x_3 and the PPF. Under this policy Country I achieves the maximum attainable degree of "exploitation" of Country II. This situation is a Paretian optimum, since the marginal rates of substitution and transformation are the same in both countries. To the extent that the optimal lump-sum royalty can be implemented, it is preferred to the three previously mentioned policies from the point of view of Country I.

If, as a result of trade with the optimal tariff imposed, Country I does become specialized in the production of *D*, the optimal royalty rate on a per-unit basis (or its equivalent of opening a branch abroad) is superior, from the point of view of Country I, to the optimal tariff. In figure 3 the optimal tariff leads Country I to consume at z_1 and specialize in producing

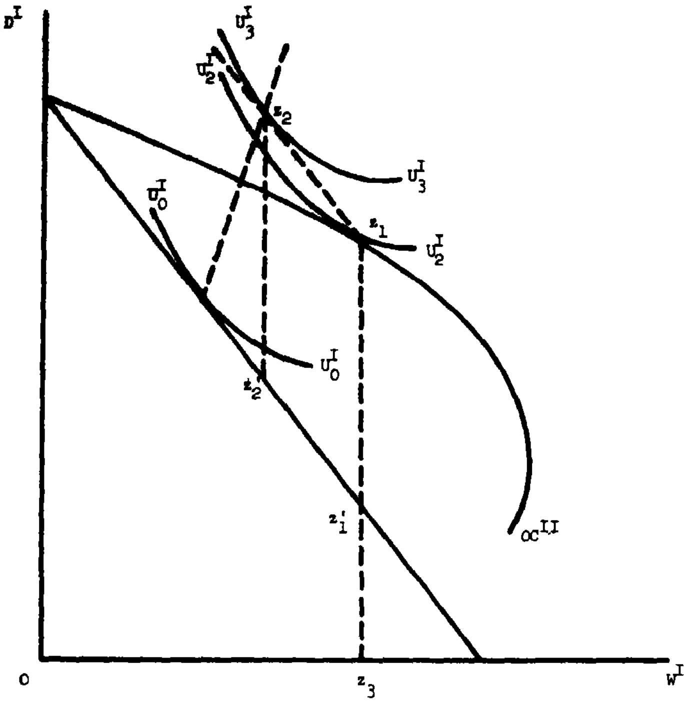

FIG. 3

D. If specialization occurs, the slope of the indifference curve at z_1 (the demand price for *D*) is larger than (or equal to) the slope of the production possibility frontier. If the technology is instead licensed to Country II with a royalty at the same rate as the optimal tariff, Country I could still consume at z_1 as under the tariff; in this case the distance z_1z_1' would be the royalty payments and the country would produce at z_1'. But in this situation, the demand price for *D* would exceed the supply price, and it would pay producers to move along the PPF to z_2', permitting consumption at z_2 and the higher level of utility U_3^1. Royalty payments account for the difference $z_2z_2' = z_1z_1'$. Similar reasoning can be used to show that licensing on a per-unit basis is always preferred to the tariff if the (constant) cost of producing *D* in the foreign country is lower than at home, and the reverse if the (constant) cost of producing *D* is higher abroad.

II

In Section I we saw that under identical constant costs and incomplete specialization in the home country it makes no difference to the home country whether it licenses foreign producers at an optimal royalty rate or trades with the foreign country and charges an optimal tariff. Since the optimal royalty rate r^* is equal to the optimal tariff t^*, it is clear that once licensing is established there are no incentives to trade, since for each unit traded the country gains t^* and loses r^*, with a zero net gain. But under increasing costs, the act of licensing per se increases world real income by allowing a more efficient utilization of world resources. Thus, it pays the home country to license foreign producers while setting the royalty and tariff rates so as to capture not only the gains which would be obtained under constant costs but also the largest possible fraction of the increase in world real income. This policy, as we shall see, implies the full equalization of marginal costs and marginal revenues from operations abroad, and, to the extent that there are increasing costs, it requires the use of both a tariff and a royalty but at different rates.

An interesting result which follows is that, if the optimal policy requires home country imports of the good whose technology is being licensed to foreign producers, the home country must apply an *import subsidy*. The reason is that the supply price of foreign producers includes a royalty rate which forces them to behave as monopolists/monopsonists in their own market. From the point of view of the home country, however, the good should be imported whenever its domestic price exceeds foreign marginal factor cost; thus, foreign producers should be subsidized to make their supply price to the home country equal to their marginal factor cost. This operation does not imply a loss to the home country, since the subsidy paid on each unit exported is recovered through the royalty paid on its production. The result is to make the price of these imports to home-country consumers equal to marginal factor cost, while still forcing foreign producers to behave as monopolists in their domestic sales.

In what follows I assume that the technology is licensed to perfectly competitive foreign producers at a given royalty rate per unit produced. Competition is also assumed to prevail among domestic producers. The home-country government can secure this competitive behavior by means of a subsidy on output (to remove the effects of any monopsony power) and a subsidy on consumption (to remove the effects of any monopoly power). If the good is traded between countries, the government of the home country can also apply a tariff.

1. *Notation*

In what follows, unstarred variables denote the home country; starred variables refer to the foreign country. The home country owns the technology to produce good 1 and licenses the technology to foreign producers at the royalty rate r.

D_i, D_i^* = consumption of good i, $i = 1, 2$.
Q_i, Q_i^* = production of good i.
p, p^* = relative market price of good 1 in terms of good 2.
c^* = relative factor cost of production of good 1 in terms of good 2 abroad (the slope of the foreign production possibility frontier). For the home country, $c = p$.
y, y^* = value of output in terms of good 2, at factor cost, where $y = pQ_1 + Q_2$ and $y^* = c^*Q_1^* + Q_2^*$.
r = royalty rate per unit of production of Q_1^* paid by foreign producers to the home country (as a fraction of factor cost).
e, e^* = private expenditures in terms of good 2.
t = export tariff (or import subsidy) imposed by the home country.
X, X^* = excess supplies of good 1.
I = real income in the home country, in terms of good 2.

2. *The Supply Side*

I adopt the standard assumptions of the neoclassical two-good two-factor model (see, e.g., Jones 1965). Factor endowments are given and not transferable between countries.

Given the tariff t and royalty rate r, domestic and foreign prices are related by

$$p(1 + t) = p^* \tag{1}$$

and

$$c^*(1 + r) = p^*. \tag{2}$$

The value of output depends on the relative factor cost of production:

$$y = y(p). \tag{3}$$

Since $p\, dQ_1/dp + dQ_2/dp = 0$ (see Jones 1965), $dy/dp = Q_1$.

$$y^* = y^*(c^*), \qquad dy^*/dc^* = Q_1^*. \tag{4}$$

The output of each good depends on the relative factor cost; in particular,

$$Q_1 = Q_1(p), \quad Q_1^* = Q_1^*(c^*), \quad dQ_1/dp > 0, \quad dQ_1^*/dc^* > 0. \tag{5}$$

3. *The Demand Side*

Demand for each good depends on the relative price and the level of private expenditures:

$$D_i = D_i(p, e) \qquad D_i^* = D_i^*(p^*, e^*). \tag{6}$$

As usual, the effect of a change in price can be divided into an income and a substitution effect:

$$\partial D_i/\partial p = S_{i1} - m_i D_1 \qquad \partial D_i^*/\partial p^* = S_{i1}^* - m_i^* D_1^*,$$

where S_{i1} and S_{i1}^* are the Hicksian (utility constant) substitution effects and m_i and m_i^* are the marginal propensities to consume good i. Demands must satisfy the budget constraints:

$$pD_1 + D_2 = e, \qquad p^*D_1^* + D_2^* = e^*, \tag{7}$$

from which it follows that

$$m_1 + m_2 = 1 \qquad m_1^* + m_2^* = 1.$$

Private expenditure abroad equals the value of output at factor cost:

$$e^* = y^*(c^*). \tag{8}$$

Private expenditures at home equal the value of output at factor cost plus the redistribution of the royalties and tariff proceeds:

$$e = y(p) + rc^*Q_1^* + tpX. \tag{9}$$

Given the budget constraints, zero world excess supply of either good is sufficient for world equilibrium in both markets. Thus, in terms of good 1, the world equilibrium condition is

$$X + X^* = 0. \tag{10}$$

The first-order change in real income for the home country (in terms of the second good) is

$$dI = p\,dD_1 + dD_2. \tag{11}$$

Substituting for dD_1 and dD_2, and recalling that $S_{11}p + S_{21} = 0$ (from a well-known property of the substitution terms), we get

$$dI = de - D_1\,dp. \tag{12}$$

Thus, when the real income of the home country is at a maximum,

$$de = D_1\,dp. \tag{13}$$

We proceed as follows. First, we derive expressions for X and X^* as functions of p, t, and r. These expressions, together with the world equilibrium condition (10), tell us how p changes in response to changes in r and t; this, together with an expression for the change is e as a function of changes in p, t, and r, allows us to express the change in the real income of the home country, from (12), as a function of changes in t and r only. Then, in the neighborhood of values of t and r which satisfy $dI = 0$ (a necessary condition for an interior maximum in real income), we have

$$\partial I/\partial r = 0, \qquad \partial I/\partial t = 0.$$

These two equations in r and t are then solved for the optimum values r^* and t^*.

4. *The Change in the Equilibrium Terms of Trade*

The first-order change in the domestic excess supply of good 1 is obtained from differentiation of

$$X = Q_1(p) - D_1(p, e),$$

which, after making use of (13), becomes

$$dX = z\, dp/p, \tag{14}$$

where $z = Q_1E - D_1N'$, the excess supply elasticity of good 1; $E = (p/Q_1)(dQ_1/dp)$, the elasticity of supply of good 1; and $N' = (p/D_1)S_{11}$, the compensated elasticity of demand for good 1.

Since we are not maximizing the foreign country's real income, we cannot assume that X^* will be free of income effects. The first-order change in X^* is

$$dX^* = dQ_1^*/dc^*\, dc^* - \partial D_1^*/\partial p^*\, dp^* - \partial D_1^*/\partial e^*\, de^*,$$

which, after making use of (1), (2), (4), and (8), becomes

$$dX^* = z^*\, dp/p + z^*\hat{t}\, dt/t - \hat{r}(z^* + D_1^*N^*)\, dr/r, \tag{15}$$

where $z^* = Q_1^*E^* - D_1^*N^* + \hat{r}m_1^*Q_1^*$, $\hat{t} = t/(1 + t)$, $\hat{r} = r/(1 + r)$, $E^* = (c^*/Q_1^*)\, dQ_1^*/dc^*$, and $N^* = (p^*/D_1^*)\, \partial D_1^*/\partial p^*$, the uncompensated demand elasticity for good 1. The change in the equilibrium p can be obtained from (14), (15), and the world equilibrium condition (10):

$$\begin{aligned} dp/p = {} & [\hat{r}(z^* + D_1^*N^*)/(z + z^*)]\, dr/r \\ & - [\hat{t}z^*/(z + z^*)]\, dt/t. \end{aligned} \tag{16}$$

The expression for the change in domestic expenditure can be obtained from differentiation of (9):

$$\begin{aligned} de = {} & [pQ_1 + \hat{r}p^*Q_1^*(1 + E^*) + \hat{t}p^*z + \hat{t}p^*X]\, dp/p \\ & + [X + \hat{r}Q_1^*(1 + E^*)]\hat{t}p^*\, dt/t \\ & + [1 - \hat{r}(1 + E^*)]\hat{r}p^*Q_1^*\, dr/r. \end{aligned} \tag{17}$$

Finally, substituting (16) and (17) into (12), we get the final expression for the change in the real income of the home country:

$$dI = (\partial I/\partial t)\, dt + (\partial I/\partial r)\, dr, \tag{18}$$

where

$$\partial I/\partial t = p[X + rQ_1^*(1 + E^*)] - \frac{z^*p}{(z + z^*)}[X + zt + rQ_1^*(1 + E^*)],$$

$$\partial I/\partial r = c^*[X + tz + rQ_1^*(1 + E^*)]\frac{(z^* + D_1^*N^*)}{(z + z^*)}$$
$$+ c^*Q_1^*[1 - r(1 + E^*)].$$

Equating these two partial derivatives to zero and solving simultaneously, we obtain the optimum values for royalty and tariff rates:

$$t^* = -1/(1 + N^*), \tag{19}$$

$$r^* = (N^* + m_1^* - E^*)/E^*(1 + N^*). \tag{20}$$

It should be noted that the optimal tariff depends only on the foreign elasticity of demand for *consumption* of good 1 and not on the elasticity of demand for imports $(-z^*/X^*)$. The optimum royalty depends on both the supply and demand elasticities. These two expressions can be given an intuitive interpretation. By the simultaneous use of the two instruments, r and t, the home country has the power to tax at different rates both foreign producers and consumers of good 1. Under these circumstances, it pays to sell to foreign consumers until the domestic price equals the foreign marginal revenue, or $p = p^*(1 + 1/N^*)$, which implies a tariff equal to that given in (19). Similarly, it pays to buy from foreign producers until domestic price equals marginal cost of purchases from them, or $p = c^*(1 + 1/E^*)$. These policies, used together, give $c^*(1 + 1/E^*) = p^*(1 + 1/N^*)$, which implies a percentage excess of p^* over c^* equal to $(N^* - E^*)/(1 + N^*)E^*$. This expression is identical with the royalty rate indicated by (20), except for the term m_1^* in the numerator, which reflects the existence of income effects associated with the use of monopsony power in the foreign country. (This effect will be analyzed in more detail in Section III.)

We can thus conclude that use of optimal tariff and royalty rates implies full maximization of profits from operations abroad. This is done by equating domestic price to marginal revenue from foreign sales and to marginal cost for foreign purchases. It is worth emphasizing the interesting implications of the optimal tariff policy which were advanced at the beginning of this section. From equation (1) it follows that if good 1 ends up being imported by the home country, t will be an import *subsidy*; otherwise t is an export tax. This guarantees that domestic price equals marginal cost of producing abroad, while still forcing foreign producers to behave as monopolists with respect to their sales in their own country.

III

In this section, I compare the optimal behavior of the home country (obtained in the last section) with that of a firm which can produce and sell in two countries at different prices. (For an analysis of the optimal behavior of the international firm in a partial-equilibrium context, see, e.g., Horst 1971). The amounts produced and sold in each country need not be equal, since output produced in one country may be sold in the other. Within a partial-equilibrium context, with no income effects shifting the demand curves, it is clear that the firm should equalize marginal costs and marginal revenues across countries in order to maximize profits. In a general-equilibrium context, however, any reallocation of production between countries brings about changes in the real value of expenditures, measured at factor costs, in each country, and this will affect the demand and marginal revenue schedules. Thus, to the extent that the marginal propensities to consume the monopolized good are nonzero, the gains from reallocating production when marginal costs are different may be more than offset by losses due to changes in demand conditions. This can be verified by maximizing the profits of a monopolist/monopsonist producing and selling a good in two countries. The same two goods are produced in each country: good 1 (produced by the monopolist) and good 2. The monopolist maximizes and collects his profits in terms of good 2. All variables retain the same meaning as in Section II.

The problem for the firm is to maximize

$$pD(p, e) + p^*D^*(p^*, e^*) - cQ(c) - c^*Q^*(c^*),$$

subject to

$$D(p, e) + D^*(p^*, e^*) = Q(c) + Q^*(c^*)$$

$$e = e(c),$$

$$e^* = e^*(c^*);$$

$$de/dc = Q(c),$$

$$de^*/dc^* = Q^*(c^*).$$

The first-order maximization conditions give

$$p(1 + N)/N = p^*(1 + N^*)/N^*, \tag{21}$$

$$c(1 + E)/E + mc/EN = c^*(1 + E^*)/E^* + m^*c^*/E^*N^*, \tag{22}$$

$$p(1 + N)/N = c(1 + E)/E + mc/EN, \tag{23}$$

$$p^*(1 + N^*)/N^* = c^*(1 + E^*)/E^* + m^*c^*/E^*N^*. \tag{24}$$

Equation (21) tells us that the marginal revenues should be equalized between countries. Equations (23) and (24) indicate that the marginal revenues in each country are equated to the direct marginal factor costs (from the point of view of the monopolist) plus an additional negative term, due to the existence of income effects. Finally, in equation (22) we can see that the direct marginal factor costs of production are not equalized across countries to the extent that the marginal propensities to consume the good are nonzero and the supply elasticities are finite.

The terms mc/EN and m^*c^*/E^*N^* can be interpreted as follows. Since the firm is exploiting its monopsony power in each country, an increase in production in one country will represent an increase in the real income of residents of that country. Suppose that the firm wants to produce an extra unit in one country for export to the other country. The quantity sold in the first country is unchanged, but its value changes, reflecting the demand curve shift due to the income effect mentioned above. Algebraically, the change in demand is given by

$$dD/dQ = (\partial D/\partial p)\, dp/dQ + (\partial D/\partial e)\, de/dQ = (ND/p)\, dp/dQ + cm/pE,$$

where use has been made of the fact that

$$de/dQ = de/dc \cdot dc/dQ = c/E;$$

but since the extra output will not be sold in that country, $dD/dQ = 0$, from which

$$dp/dQ = -cm/DEN.$$

Thus, the value of sales in that country changes by

$$D\, dp/dQ = -cm/EN.$$

The total full costs of that operation are then

$$d(cQ)/dQ - D\, dp/dQ = c(1 + E)/E + cm/EN.$$

Since profit is maximized, this must be equated to the marginal revenue from selling the extra unit abroad, $p^*(1 + N^*)/N^*$. Thus, profit maximization requires that

$$p^*(1 + N^*)/N^* = c(1 + E)/E + cm/EN,$$

which is a result that follows from (21) and (23) above. The other optimality conditions can be explained by similar reasoning.

Assume now that we are interested in maximizing the real income of the home country of the international firm (taken to be the one represented by the unstarred variables). The real income of the home country will be maximized when the international firm behaves as a competitor at home and fully maximizes its profits from operations abroad. Competitive behavior at home implies that the marginal cost of production (as seen by the firm) must be equal to the marginal rate transformation c, and that the marginal revenue from domestic sales (as seen by the firm) must be equal to the domestic marginal rate of substitution p. Both results can be obtained by means of a subsidy on domestic production plus a subsidy on domestic consumption. In this case, the conditions yielding maximum profit for the firm (and maximum real income for the home country) are

$$p = p^*(1 + N^*)/N^*, \tag{25}$$

$$c = c^*(1 + E^*)/E^* + m^*c^*/E^*N^*, \tag{26}$$

$$c = p, \tag{27}$$

$$p^*(1 + N^*)/N^* = c^*(1 + E^*)/E^* + m^*c^*/E^*N^*, \tag{28}$$

which are exactly identical with those obtained in Section II under licensing to competitive foreign producers. To the extent that the international firm is able to charge a higher price abroad than at home, as required in (25), the government of the home country need only force the firm to behave competitively in the home markets. But if price discrimination is not possible, there are two alternative ways to attain the optimum. (*a*) The government can grant an exclusive export license to the international firm. This will automatically give the firm the power to charge the required higher price abroad. (*b*) The government can impose an export tariff at the rate $-1/(1 + N^*)$, from which the international firm is exempted. This is essentially the same as (*a*), since under those conditions only the international firm will be in a position to export. If the firm is not exempted, the optimal solution will not be attained when, in equilibrium, the domestic plant of the international firm is exporting to the foreign country. This is because the firm has two ways to supply the foreign market: by domestic production at a marginal cost of $c(1 + t)$, where t is the export tariff; or by foreign production at a marginal cost of $c^*(1 + E^*)/E^* + m^*c^*/E^*N^*$. In maximizing profits, the firm will make both marginal costs equal, and thus condition (26) will be violated. (If the home country ends up importing from abroad, the tariff will not affect the operations of the firm and will simply avoid the possibility of arbitrage by consumers.)

Conclusions

We have developed, within the context of a two-country general-equilibrium model, the optimal trade and licensing policies for a country which is the only owner of the technology to produce one good. The optimal behavior turns out to entail (*a*) the full exploitation of any monopoly and/or monopsony power that the owner of the technology may possess in the foreign market for his product and in the foreign market for the factors used in its production, and (*b*) competitive behavior with respect to the use of domestic resources and the sale of the product in the domestic market. In the case of licensing competitive foreign producers, the optimal tariff and royalty rate obtained secure this result by allowing the equalization of the domestic price with both the marginal revenue from foreign sales and the marginal factor cost (from the point of view of the home country) of producing an extra unit abroad. If production is carried out in both countries by a single domestic firm, the government of the home country should induce the firm to behave competitively at home while allowing it to fully maximize its profit from operations abroad.

References

Horst, Thomas. "The Theory of the Multinational Firm: Optimal Behavior under Different Tariff and Tax Rates." *J.P.E.* 79, no. 5 (September/October 1971): 1059–72.

Jones, Ronald. "The Structure of Simple General Equilibrium Models." *J.P.E.* 72, no. 6 (December 1965): 557–72.

———. "International Capital Movements and the Theory of Tariffs and Trade." *Q.J.E.* 81, no. 1 (February 1967): 1–38.

Kemp, Murray. "The Gains from International Trade and Investment: A Neo Heckscher-Ohlin Approach." *A.E.R.* 56, no. 4 (September 1966): 788–809.

Full-Time Schooling in Life-Cycle Models of Human Capital Accumulation

---

T. D. Wallace and L. A. Ihnen

Duke University and North Carolina State University

A reduced-form equation relating length of "formal schooling" to market, endowment, and ability parameters was derived for a life-cycle human capital accumulation model with alternative assumptions: (*a*) equal borrowing and lending rates and (*b*) no loans for human capital investment. Length of "formal schooling" increases when loans are unavailable. For both cases, length of "formal schooling" varies directly with length of work life, a Hicks-neutral "ability" index, and the ratio of the human capital rental rate to the price of associated inputs, and varies inversely with the discount rate, deterioration rate, and initial human capital stock.

I. Introduction

The Ben-Porath (1967) model of human capital accumulation is an imaginative attempt to give a neoclassical basis for inference about investments in human capital. He assumed that the individual was confronted with exogenous economic forces such as the price of educational inputs, a rental rate on human capital, and a market rate of interest for borrowing and lending for whatever purpose, and fixed retirement time. Further, he assumed internal constraints: (i) a fixed deterioration rate on the stock of human capital, (ii) an initial exogenous stock on which to build, (iii) an internalized production function whose parameters and form determine the ability to augment the stock of human capital, and (iv) a recognition that human capital as an input to produce future human capital resides in the individual.

Journal Series Paper 4308, North Carolina State University Agricultural Experiment Station. We received valuable comments on a draft of this work from T. W. Schultz, Yoram Ben-Porath, Mary Jean Bowman, Lee Lillard, William Haley, and Thomas Johnson. The research was supported in part by National Institute of Education grant NE-G-00-3-0152.

[*Journal of Political Economy*, 1975, vol. 83, no. 1]

© 1975 by The University of Chicago. All rights reserved.

The Ben-Porath solution yielded the result that early in life the individual specializes in the production of human capital. Thus, his solution broke into two phases, the first phase characterized by zero earnings. If one therefore directs attention to the shape of earnings streams implied by the model, the second phase deserves greater attention; Ben-Porath concentrated his attention on the second phase. In fact, Ben-Porath did not solve his model for some of the reduced forms implicit in it. In particular, he did not obtain a formal expression relating the end of the period of specialization to the parameters of the model. It is of some importance to focus attention on implications of a Ben-Porath-type model regarding the length of the period of specialization in the production of human capital. That is, much of the empirical work on earnings takes the length of formal schooling as a datum, not accounting for the possibility that the individual is free to decide how long to stay in school, depending on his environment, his ability, and his market alternatives. Elsewhere, time of labor force entry and length of time spent in school are taken as the variables to be explained, but often such studies are not guided by any assumed formal choice mechanism.

One of the objectives of the present paper is to give a solution for the path of human capital accumulation during the "formal schooling" phase of the Ben-Porath model and use that solution to characterize the end of the period of specialization in terms of the exogenous parameters. Another objective is to develop and examine a similar model given an alternative assumption that there are no opportunities for borrowing for the purpose of investing in human capital.

The main justification for our second objective is pragmatic. That is, the reduced form relating the length of formal schooling to the parameters of the model turns out to be more tractable in our case than in Ben-Porath's, and a numerical experiment shows that, over at least some range of the parameter space, directional effects of parameter changes on the length of formal schooling are invariant between the two models. Moreover, some general analytical results regarding the Ben-Porath period of specialization become apparent through a comparison of the reduced forms implicit in the two models. Thus, the work reported here provides a basis for inference and a guide to empirical work regarding the length of formal schooling.

Our second objective is partially justified by the common supposition running through the literature on human capital that loan markets for investment in education are imperfect.[1] Many such speculations can be cited, and we have chosen the following examples.

---

[1] In private correspondence, Ben-Porath pointed out that his assumption about loan rates does not preclude the exogenous rates from varying for different individuals. However, in this paper, we define "imperfection" to be a divergence in the borrowing rate for financing educational investment and other rates.

Schultz (1961, p. 4) states: "It is indeed elementary to stress the greater imperfection of the capital market in providing funds for investment in human beings than for investment in physical capital." Concurring, Friedman (1962, p.102) states: "Investment in human capital cannot be financed on the same terms or with the same ease as investment in physical capital." Indeed, for some students it is held that the loan price is so high that one can in effect say that loan funds do not exist. Thus, Friedman and Kuznets (1954, p. vi) in an earlier work state: "The economic stratification of the population is important because capital invested in professional training, unlike capital invested in factories and machines, can rarely be obtained in the open market; it must be provided by the prospective practitioner himself, his parents, or a benefactor."

If no loans are available to the individual for educational investment, the separation theorem can be retained only by artifact. That is, to retain equivalence of maximizing present value of income and utility, one must now assume that borrowing rates for individual consumption (or other types of investment) are exogenous and equal to the lending rate. As Becker (1964) points out, such an assumption is a bit specious, because the borrower has some freedom to convert funds to alternative uses. However, much consumer borrowing is for durables (autos, houses, appliances) and the lender frequently retains title, thus precluding conversion. Whatever the value of these remarks, we proceed by maximizing the present value of the rental flow of human capital, and, as a consequence of our "no educational loans" assumption, we have less cause for invoking the separation theorem than did Ben-Porath.

We agree with a reviewer of an earlier draft of this paper that retention of the separation theorem with divergence of lending-borrowing rates is a luxury. And we agree with another reviewer that zero borrowing for schooling is not descriptive. The "truth" lies somewhere between, and choice thus depends on individual time preference. Since our assumptions lead to an implication that the length of schooling is greater for every case than the length of schooling in the Ben-Porath model, bias runs in the same direction for both models. That is, divergence in borrowing-lending rates that dictates choice via intertemporal utility leads to the possibility that some individuals would fail to invest, assuming, of course, that the borrowing rate exceeds the lending rate rather than vice versa.

The objectives we attempt to achieve in this paper should be viewed as intermediate steps en route to confronting these models with data. To date, most empirical effort guided by the Ben-Porath-type model has been directed toward earnings data.[2] The solutions given here and the resulting reduced form equations have implications for alternative data.

---

[2] See Ben-Porath (1970), Haley (1971), and Lillard (1972).

II. A Ben-Porath-Type Model with No Loans for Educational Investment

What follows is a presentation and analysis of a human capital accumulation model similar to Ben-Porath's except for the loan market assumption. For expository purposes, it is necessary to duplicate some of the Ben-Porath discussion.

An individual is assumed to arrive at a specified age, t_0, with an acquired and exogenous stock of human capital, E_0. The question of units of measurement for E_t, the stock of human capital at time t, is begged. It is a stock of "*Eds*" from which rental earnings may flow. Potential earnings are

$$Y_t^* = RE_t, \qquad t \geq t_0, \tag{1}$$

and it is assumed that the individual has acquired no assets at initial time other than his stock of human capital. The rental rate, R, is fixed.

The individual can add to his human capital stock, thereby augmenting potential earnings. Assume that the gross addition to "*Eds*" is technically constrained by a Cobb-Douglas production function,

$$Q_t = \beta_0 K_t^{\beta_1} D_t^{\beta_2}, \tag{2}$$

where β_0, β_1, and β_2 are positive and scale (i.e., $\beta_1 + \beta_2$) is less than unity.[3]

The symbol Q_t represents the gross addition to human capital in time t; K_t is that part of E_t used to augment human capital as opposed to earning an immediate cash flow; and D_t is other educational inputs taken as a bundle.

The price of educational inputs is taken to be exogenous and fixed through time. The price of a unit of D is denoted by the symbol P. The opportunity cost of K is the rental rate, R. Hence, net earnings is potential earnings net of investment in schooling. That is, net earnings, Y_t, is

$$Y_t = R(E_t - K_t) - PD_t, \qquad t \geq t_0. \tag{3}$$

We assume that educational inputs must be purchased out of current cash flow. That is, contrary to the Ben-Porath assumption that the individual can defray outlay for tuition, books, etc., by borrowing, we assume that net earnings are always nonnegative. Thus, a constraint that must be met is

---

[3] In Ben-Porath's notation, our E_t is his K_t and our K_t is his $s_t K_t$; that is, Ben-Porath's s_t is our K_t/E_t. As noted by a reviewer, some of our results do not depend on this specific form for the production process. The nature of this invariance is clarified at the appropriate points.

$$Y_t = g^2(t), \qquad t \geq t_0, \tag{4}$$

where $g(t)$ is any arbitrary differentiable real time function.[4]

Note that (4) implies $E_t \geq K_t$, since R, P, and D_t are nonnegative.[5] Assuming that human capital deteriorates at a constant rate in the absence of augmentation, the net addition to the human capital stock is governed by

$$\dot{E}_t = Q_t - \delta E_t, \qquad 0 < \delta < 1, \tag{5}$$

where the dot notation indicates a time derivative and δ is the rate of deterioration of the human capital stock.

It is assumed that there is a fixed endpoint in time, t_n, at which the individual retires. This implies that R, the rental rate on human capital, jumps to zero at time t_n.

Finally, it is assumed that the individual faces an exogenous market rate of interest, r, at which he can loan or borrow for consumption purposes, and the objective is assumed to be that of maximizing the present value of net earnings, given the restrictions as stated. That is, the objective is to maximize

$$J = \int_{t_0}^{t_n} \Lambda \, dt, \tag{6}$$

where

$$\begin{aligned} \Lambda = e^{-rt}[R(E_t - K_t) - PD_t] - \lambda_{1t}(\dot{E}_t - \beta_0 K_t^{\beta_1} D_t^{\beta_2} + \delta E_t) \\ - \lambda_{2t}[R(E_t - K_t) - PD_t - g^2(t)] \end{aligned} \tag{7}$$

and λ_{1t} and λ_{2t} are Lagrangians.[6]

Before we present results, table 1 is given so that the reader may have a ready list of the variables of the model.

A. Necessary Conditions for Maximizing J

The reader is referred to Sage (1968) for a general discussion of necessary conditions for maximizing an integral over time and to Haley (1971) for

---

[4] The arbitrary real function $g(t)$ enters the model as an endogenous variable. By making net earnings equal to its square, we ensure the nonnegativity of net earnings (see Sage 1968).

[5] The Ben-Porath assumption analogous to (4) is that human capital diverted to production cannot be greater than the total stock available.

[6] See Sage (1968) for methods of incorporating equality constraints into variational calculus problems.

TABLE 1

LIST OF VARIABLES: MODEL FOR HUMAN CAPITAL INVESTMENT WITH NO LOANS

ENDOGENOUS		EXOGENOUS	
Variable	Description	Variable	Description
Y_t^*......	Potential earnings	R.......	Rental rate on human capital
Y_t......	Net earnings	δ.......	Rate of deterioration of E_t
E_t......	Human capital stock	P.......	Price of educational inputs
K_t	Part of E_t being reinvested	$\beta_0, \beta_1, \beta_2$	Production function parameters
D_t	Purchased educational inputs	E_0......	Initial stock of human capital
Q_t	Produced human capital		
$\dot{E}_t$......	Net additions to E_t	t_0, t_n	Initial and ending times of planning horizon†
λ_{1t}......	Shadow price of Q_t^*		
λ_{2t}......	Part of shadow price of D_t, K_t^*	r	Market rate of interest‡
$g(t)$.....	Square root of net earnings		
t^*	Point in time at which net earnings turn positive*		

* Implicit in solution.
† t_0 is taken to be zero without loss in generality.
‡ At which loans or borrowing for consumption purposes can take place.

a complete discussion in a similar case.[7] In the present case, the general necessary conditions are:

$$Re^{-rt} - \delta\lambda_{1t} - R\lambda_{2t} + \dot{\lambda}_{1t} = 0, \tag{8}$$

$$\lambda_{1t}\beta_0\beta_1 K_t^{\beta_1 - 1} D_t^{\beta_2} = R(e^{-rt} - \lambda_{2t}), \tag{9}$$

$$\lambda_{1t}\beta_0\beta_2 K_t^{\beta_1} D_t^{\beta_2 - 1} = P(e^{-rt} - \lambda_{2t}), \tag{10}$$

$$\dot{E}_t = \beta_0 K_t^{\beta_1} D_t^{\beta_2} - \delta E_t, \tag{11}$$

$$R(E_t - K_t) - PD_t = g^2(t), \tag{12}$$

$$2g(t)\lambda_{2t} = 0. \tag{13}$$

Note that equations (11) and (12) are simply restatements of the constraints. Equations (9) and (10) are "value of marginal product equals factor price" equations. Equation (8) is a differential equation describing the path of the implicit price of gross additions to human

---

[7] In general, if the objective is to maximize

$$\int_{t_0}^{t_n} \Lambda(\mathbf{x}, \dot{\mathbf{x}}, t)dt,$$

the Euler-Lagrange conditions are

$$\frac{\partial\Lambda}{\partial\mathbf{x}} - \frac{d}{dt}\frac{\partial\Lambda}{\partial\dot{\mathbf{x}}} = 0,$$

where $\mathbf{x}$ is a vector of time variables, $\dot{\mathbf{x}}$ is the vector of time derivatives, etc. To these must be added transversality conditions, but for the present problem it becomes clear how to resolve constants of integration without appeal to transversality conditions.

capital (price of Q_t), and equation (13) says that if the shadow part of the input prices is zero, net earnings are not necessarily zero, or, conversely, if net earnings are zero, there can be a shadow component to input prices. This makes economic sense because when the individual specializes in educational investment (zero net cash flow), his human capital stock is a fixed and limiting input as of a moment in time.

Thus, the solution can break into two phases, and, where it exists, we denote the period of zero net earnings as Phase I, the period of "formal schooling."[8]

B. Phase I

Note that even when λ_{2t} is nonzero, as it is in Phase I, a ratio of the two *VMP* curves, equations (9) and (10), eliminates the shadow part of the input prices. This is in contrast to the Ben-Porath case and makes for a simpler solution.[9] Hence, the important relationships for characterizing the period of "formal schooling" are

$$\frac{\beta_1 D_t}{\beta_2 K_t} = \frac{R}{P}, \tag{14}$$

$$E_t = K_t + \frac{P}{R} D_t, \tag{15}$$

$$\dot{E}_t = \beta_0 K_t^{\beta_1} D_t^{\beta_2} - \delta E_t. \tag{16}$$

Equations (14) and (15) imply

$$K_t = \frac{\beta_1}{\beta_1 + \beta_2} E_t, \tag{17}$$

$$D_t = \frac{\beta_2}{\beta_1 + \beta_2}\left(\frac{R}{P} E_t\right). \tag{18}$$

Substituting (17) and (18) into (16), we get a differential equation in E_t alone:

$$\dot{E}_t - \mu E_t^{1-\Delta} + \delta E_t = 0, \tag{19}$$

---

[8] The conditions under which Phase I exists become clear in the derivation of the optimal path of human capital stock, E_t, which follows. Note that when loans are not available, the end of the period of specialization or "formal schooling" is not the time at which labor force entry occurs but instead is the time at which net earnings become positive, that is, completion of "full-time schooling." This result is consistent with observed behavior, for as one reviewer noted entry into the labor force often precedes completion of schooling.

[9] See Appendix.

where

$$\mu = \beta_0 \left(\frac{\beta_1}{1 - \Delta}\right)^{1-\Delta} \left(\frac{R\beta_2}{P\beta_1}\right)^{\beta_2} \quad \text{and} \quad \Delta = 1 - \beta_1 - \beta_2.$$[10]

An explicit solution for E_t from (19) is contained in the following equation:

$$-\frac{1}{\delta\Delta} \log (\mu - \delta E_t^{\Delta}) = t + c, \tag{20}$$

where c is a constant of integration.

The constant, c, can be found by setting $t = t_0$ in (20). Without loss of generality we take $t_0 = 0$, so

$$c = -\frac{1}{\delta\Delta} \log (\mu - \delta E_0^{\Delta}). \tag{21}$$

Resolution of the constant of integration instructs us that Phase I exists only if μ is larger than δE_0^{Δ}. Otherwise c is undefined.[11]

If we incorporate the constant of integration and rewrite (20) in altered form, the optimal path of human capital stock during specialization is described by

$$E_t^I = \left[\frac{\mu}{\delta} + \left(E_0^{\Delta} - \frac{\mu}{\delta}\right) e^{-\delta\Delta t}\right]^{1/\Delta}, \qquad 0 \leq t \leq t^*, \tag{22}$$

---

[10] As pointed out by a reviewer, a similar differential equation governing accumulation of "*Eds*" in Phase I would be obtained from any homogeneous production function. The parameter μ would remain a function of relative prices—only the parametric representation would vary. For example, if one chose a constant elasticity of substitution (CES) function,

$$Q_t = (\alpha_1 K_t^{-\beta} + \alpha_2 D_t^{-\beta})^{-v/\beta},$$

where v is scale (less than unity), then equation (19) would become

$$\dot{E}_t - \tilde{\mu} E_t^{v} + \delta E_t = 0,$$

where

$$\tilde{\mu} = \left[\alpha_1 + \alpha_2 \left(\frac{\alpha_2 R}{\alpha_1 P}\right)^{-\beta\sigma}\right]^{-(v/\beta)} \div \left[1 + \frac{P}{R}\left(\frac{\alpha_2 R}{\alpha_1 P}\right)^{\sigma}\right]^{v}$$

and where σ is the elasticity of substitution $1/(1 + \beta)$. A similar generality holds in Phase II, as noted by the same reviewer (see n. 14 below). The reasons for working with the Cobb-Douglas are to preserve manageable closed forms for relating the length of formal schooling to parameters and to facilitate comparison with similar closed forms implicit in the Ben-Porath model.

[11] Since the net accumulation of human capital is at the rate $Q_t - \delta E_t$, and optimal Q_t in Phase I is $\mu E_t^{1-\Delta}$, the requirement that μ be larger than δE_0^{Δ} is simply a requirement that the initial optimal gross rate of capital accumulation be larger than deterioration. Thus, Phase I, if it exists at all, is a period of positive net additions to the stock of human capital.

where the superscript on E_t^I indicates Phase I and where t^*, which marks the end of the period of specialization in the production of human capital, remains to be determined.

From (22), the general shape of human capital accumulation during Phase I can be discerned. The stock begins at E_0 and, given the existence of a specialization phase, monotonically increases to an upper asymptote, the upper asymptote being $(\mu/\delta)^{1/\Delta}$. The path of accumulation has the potential of passing through a point of inflection.

Via equations (17) and (18), similar paths can be immediately constructed for K_t and D_t in Phase I. Note that the optimal fraction of human capital devoted to production (Ben-Porath's s_t) is $\beta_1/(1 - \Delta)$ during Phase I: it is thus constant during specialization but smaller than unity.

C. *Phase II*

During the period of positive net earnings, the shadow price component of input prices, λ_{2t}, becomes zero. Hence, the relationships governing Phase II important to our purpose are

$$\lambda_{1t}\beta_0\beta_1 K_t^{\beta_1 - 1} D_t^{\beta_2} = Re^{-rt}, \tag{23}$$

$$\lambda_{1t}\beta_0\beta_2 K_t^{\beta_1} D_t^{\beta_2 - 1} = Pe^{-rt}, \tag{24}$$

$$Re^{-rt} - \delta\lambda_{1t} + \dot{\lambda}_{1t} = 0. \tag{25}$$

As previously noted, λ_{1t} is the shadow price of a unit of produced "*Eds*." Since at time t_n it is assumed that the rental rate on human capital goes to zero, it follows that the shadow price of produced "*Eds*" goes to zero at t_n.[12] This furnishes an integrating constant for the general solution to equation (25); hence, a specific solution to (25) is

$$\lambda_{1t} = \frac{R}{r + \delta} e^{-rt}[1 - e^{(r+\delta)(t - t_n)}], \qquad t^* \leq t \leq t_n.^{13} \tag{26}$$

Using equation (26) in conjunction with the two *VMP* curves, (23) and (24), the optimal time path for the part of educational stock diverted as an input during Phase II is

---

[12] In variational calculus language, educational stock is not constrained to be any particular value at final time; hence its "perturbation" is not zero at t_n, so the transversality condition implies that λ_{1t} evaluated at t_n must be zero (see Sage 1968).

[13] Equation (25) is a standard linear differential equation for which a general solution may be found in textbooks on differential equations.

$$K_t^{II} = \beta_1(1 - \Delta)\,\frac{1 - \Delta}{\Delta}\,\mu^{1/\Delta}\left[\frac{1 - e^{(r+\delta)(t-t_n)}}{r + \delta}\right]^{1/\Delta}, \qquad (27)$$

$$t^* \leq t \leq t_n.^{14}$$

An optimal path for D_t in Phase II can be similarly determined. The differential equation describing the optimal E_t^{II} path is difficult to integrate but can be handled.[15] However, it is not vital to our purposes and will be omitted from consideration here.

D. End of the Period of Specialization

At the point $t = t^*$, it must be true that the optimal stock of human capital, E_t^I, in Phase I coincides with E_t^{II} in Phase II. At the point $t = t^*$, equation (17) must also hold true. Thus there is a nondifferentiable (unsmooth) point in the E_t path at t^*, but there is no provision in the model for an instantaneous gain or loss in human capital at a single point in time. Therefore, at $t = t^*$,

$$K_{t^*} = \frac{\beta_1}{1 - \Delta}\,E_{t^*}. \qquad (28)$$

Hence, to determine t^*, the relevant functions to equate are

$$\frac{1 - \Delta}{\beta_1}\,K_{t^*}^{II} = \left[\frac{(1 - \Delta)\mu}{r + \delta}\right]^{1/\Delta}[1 - e^{(r+\delta)(t^*-t_n)}]^{1/\Delta}, \qquad (29)$$

$$E_{t^*}^{I} = \left(\frac{\mu}{\delta}\right)^{1/\Delta}\left[1 - \left(1 - \frac{\delta}{\mu}\,E_0^{\Delta}\right)e^{-\delta\Delta t^*}\right]^{1/\Delta}. \qquad (30)$$

Thus, a reduced-form equation relating the end of "formal schooling" to the parameters of the model is

$$1 - \left(1 - \frac{\delta}{\mu}\,E_0^{\Delta}\right)e^{-\delta\Delta t^*} = \frac{\delta(1 - \Delta)}{r + \delta}\,[1 - e^{(r+\delta)(t^*-t_n)}]. \qquad (31)$$

---

[14] With reference to n. 10, if the production function were CES, equation (27) would read

$$K_t = (\nu\alpha_1)^{1/(1-\nu)}\left[\alpha_1 + \alpha_2\left(\frac{\alpha_2 R}{\alpha_1 P}\right)^{\sigma}\right]^{(\nu+\rho)/(\nu\rho-\rho)}\left[\frac{1 - e^{(r+\delta)(t-t_n)}}{r + \delta}\right]^{1/(1-\nu)}.$$

This is an example of the fact that, in general, the shapes of the optimal time paths involved are similar for any homogeneous production function, with scale being crucial. Only the constants change with variations in the function. Thus, anticipating the results given in Section III, we find that the partial effects of changes in r, δ, t_n, and E_0 the length of formal schooling are invariant to the form of the production function within the class of homogeneous functions.

[15] See Haley (1972).

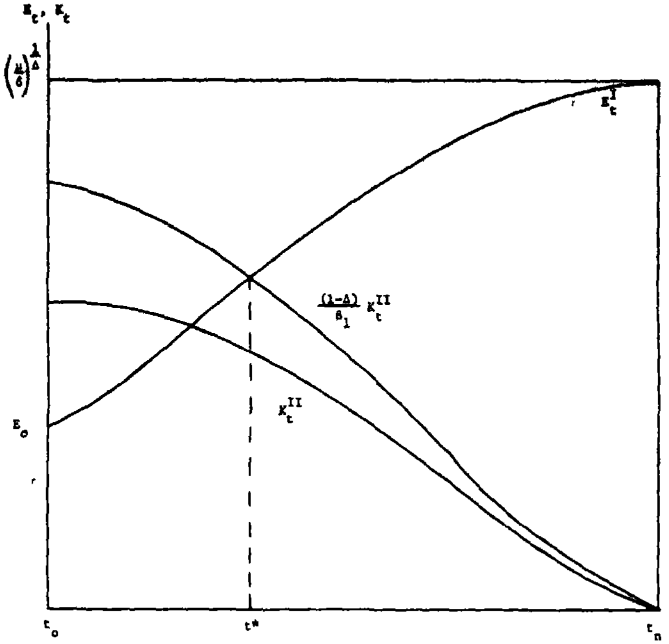

FIG. 1.—Determination of the end of "formal schooling"

The function on the left-hand side of (31) is an increasing function in t^*, and the right-hand side is a decreasing function in t^*.

Figure 1 pictures the crossing of the Phase I and Phase II functions that determine t^*.

III. Partial Effects of Parameter Changes on the Length of the Period of Specialization

Equation (31) is a reduced-form equation relating the length of the period of specialization (t^*) to the parameters of the model. Some of the partial effects of changes in the parameters on the length of specialization can be visualized with the aid of figure 1. Other effects are more difficult to visualize but can be derived analytically from the reduced form via an implicit-function theorem.[16]

An increase in the initial stock of human capital, E_0, shifts up the Phase I function of figure 1 without affecting the Phase II function. Hence, an increase in the initial stock of human capital, holding other

---

[16] The useful implicit-function theorem is that if $F(x_1, x_2, \ldots, x_n) = 0$, $\partial x_i/\partial x_j = (-\partial F/\partial x_j)/(\partial F/\partial x_i)$. In the present case,

$$F = 1 - e^{-\delta\Delta t^*} - \frac{\delta(1 - \Delta)}{r + \delta}[1 - e^{(r+\delta)(t^*-t_n)}] + \frac{\delta}{\mu}E_0^{\Delta}e^{-\delta\Delta t^*}.$$

parameters constant, decreases the length of the period of specialization in the production of human capital.[17] This result may seem inconsistent with casual observation, for one would expect to find a positive correlation between, say, parental schooling or income and length of time spent in school by the offspring. There is an explanation for this apparent anomaly. First, the effect under discussion is partial; ability (as reflected by the parameters of the production function) and the price of schooling inputs are held constant. Students with higher ability also will tend to have a larger initial stock of human capital; thus, simple and partial correlations may have opposite signs. Similarly, parents with larger resources may subsidize their child's education for longer periods, thereby lowering the price of educational inputs. In addition, since institutions of higher education use nonprice rationing, the supply of schooling to an individual is likely to be influenced by both his ability and his achievement prior to enrollment. The reduced form under discussion here is a demand function. Finally, it should be reiterated that education in these models is not treated as a consumer good—for either the individual or his family.

Again, with reference to figure 1, a lengthening of work life, t_n, has the effect of shifting the Phase II function upward without affecting the path of human capital accumulation in Phase I. Thus, a postponing of retirement has the partial effect of lengthening the period of "formal schooling."

Other partial effects are more difficult to visualize with the aid of figure 1. For example, changes in the rate of deterioration, δ, the rental-price ratio, R/P, and the production function parameters affect the functions in both phases. Thus, an increase in the rental-price ratio, R/P, shifts both functions upward. This increase in R/P clearly increases discounted future earnings, but its effect on length of schooling is less clear. However, use of the implicit-function theorem establishes the result that the partial effect of increasing the rental-price ratio increases the period of specialization.

Other partial effects that can be established analytically from equation (31) are: (i) changing the interest rate, r, affects time in school in an inverse manner. Increasing r decreases t^* and conversely. (ii) Similarly, increasing the Hicks-neutral productivity of the individual in his capacity to augment human capital (increase β_0) increases the length of specialization in the production of human capital.

Changes in the remaining parameters, δ, β_1, and β_2, are ambiguous in their effects. However, in a simulation exercise some results were established (see Section V).

IV. End of the Period of Specialization in the Ben-Porath Case

The crucial difference in the Ben-Porath case is that there is no shadow price for the D_t input in the specialization phase, since the individual is not constrained to pay as he goes for these inputs. The amount of "*Eds*" possessed by the individual is constraining during Phase I for both versions of the model; hence, the per unit cost of K_t as an input has an internal valuation. When the individual is specializing in the Ben-Porath sense, he wishes he could purchase additional "*Eds*" to use as an input but cannot, since they are internal to the individual. The end of the period of specialization is characterized by a coinciding of the shadow price of "*Eds*" and the opportunity cost of using "*Eds*" in production. Thus, in the Ben-Porath version, Phase I ends as λ_{2t} goes to zero and the individual is content to move along a usual Cobb-Douglas expansion path. In contrast, D_t is shadow priced in the same way as K_t in the no-borrowing situation; that is, all prices are internal. This difference makes the solution for Ben-Porath Phase I quite complex, and, in fact, no solution for Phase I was presented in his paper.

---

[17] As noted earlier, Phase I exists only if E_0 is less than $(\mu/\delta)^{1/\Delta}$.

Presented below (eq. [32]) is an implicit function relating the end of the period of specialization in Ben-Porath's model to the parameters. Its complexity makes for analytical difficulty; however, some general results can be obtained by comparing the Ben-Porath model with the no-loans case. Both models have the same Phase II solution for K_t, the human capital input into the production process. The Phase I path for E_t lies above and to the left of the path of human capital accumulation during the zero net earnings phase for the no-loans case, since the Ben-Porath individual has more inputs at his disposal during specialization. Hence, the period of specialization in the Ben-Porath model is shorter than the length of "formal schooling" in our model. Also, initial human capital, E_0, only affects the Phase I function; therefore, the Ben-Porath length of specialization varies inversely with E_0. In his paper, Ben-Porath (1967) recognized the effect of E_0 on length of specialization.

Equation (32) is derived in the Appendix:

$$\begin{aligned} 1 - \delta e^{[\gamma\beta_2 - \delta(1-\beta_1)]\tau} - \frac{\delta}{\theta}\beta_0\beta_1 \left[\frac{1 - e^{(r+\delta)(\tau - t_n)}}{r + \delta}\right] \\ + (\beta_0\beta_1)^{-\beta_2/\Delta}\left(\frac{\beta_1 P}{\beta_2 R}\right)^{[\beta_2(1-\beta_1)]/\Delta} E_0^{1-\beta_1} e^{-\delta(1-\beta_1)\tau} \\ \times \left[\frac{1 - e^{(r+\delta)(\tau - t_n)}}{r + \delta}\right]^{-\beta_2/\Delta} = 0, \quad (32) \end{aligned}$$

where τ is the end of Ben-Porath's period of specialization and

$$\gamma = \frac{\delta + r - \delta\beta_1}{1 - \beta_2}, \quad (33)$$

$$\theta = \frac{\beta_0(1 - \beta_1)\delta}{\gamma\beta_2 + \delta(1 - \beta_1)}. \quad (34)$$

TABLE 2

PARAMETER VALUES USED FOR SOLUTIONS FOR t^* AND τ

β_0	β_1	β_2	E_0	δ	r	R/P	t_n
30	0.2	0.3	2,000	0.06	0.07	0.75	40
40	0.25	0.25	3,000	0.07	0.09	1.0	50
50	0.3	0.2	4,000	...	...	1.25	...

Equations (31) and (32) were solved for several specific parameter sets to obtain some partial comparisons; the numerical results are given in Section V.

V. Some Numerical Comparisons

Table 2 contains the values of parameters for which equations (31) and (32) were solved for t^* and τ.[18] The following considerations were a guide to choosing the values in table 2:

i) Since only the price ratio R/P is relevant, no generality is lost in fixing R at unity. Thus, "*Eds*" can be thought of as being identical to potential earnings. In particular, the reader can think of initial stock of "*Eds*," E_0, as potential annual earnings in dollars at initial time. Units of D_t are as arbitrary as units of K_t. However, varying the price ratio R/P around unity implies a D measure in annual units roughly the same as K. For example, in Phase I, when net income is zero and R/P and β_1/β_2 are unity, then $D = K = \frac{1}{2}E$ in the no-loans model.

ii) Values for human capital deterioration were picked close to those reported in Johnson (1970), and values for r were chosen in a range that seemed reasonable for real rates of return to nonhuman capital. Both r and δ are continuous discount rates; therefore, annual discrete equivalents would be slightly larger than the r and δ values in table 2.

iii) Values of β_1 and β_2 were chosen to sum to $\frac{1}{2}$. This allows a simple solution to the differential equation governing potential and net earnings during Phase II.[19]

iv) The parameter, t_n, represents expected years of productive life. That is, if the individual is assumed to make his choice at age 18, say, then $t_n = 50$ implies retirement at age 68.

v) The parameter β_0 is chosen to crudely represent "ability." The value 40 was chosen so that the individual would realize an initial gross internal rate of return of about 50 percent when $\beta_1 = \beta_2 = \frac{1}{4}$, $R/P = 1$, and $E_0 = 3{,}000$. (Initial production, Q_0, would be about 1,500 for this set of values.) Variations above and below 40 represent variations from a "norm" of a Hicks-neutral "ability" measure.

---

[18] Richard Brook developed the solution programs.

[19] Two reviewers of an earlier draft suggested some discussion of earnings profiles implicit in the parameter sets.

Results of the numerical comparisons were as follows:

i) Length of specialization in the Ben-Porath model, τ, was uniformly smaller than the period of zero net earnings in the no-loans case (t^*). As noted previously, this result has general validity, since with a no-loan restriction the individual can increase his capital stock at a faster rate during the specialization phase.

ii) Both τ and t^* varied directly with β_0, R/P, and t_n and inversely with δ, r, and E_0. At least for the initial human capital stock (E_0), generality can be claimed for the Ben-Porath model.

iii) For the parameter sets chosen, an increase in β_1 with scale fixed has ambiguous effects on t^*, depending on the price ratio, R/P. In contrast, increases in β_1 with offsetting decreases in β_2 to fix scale increased the Ben-Porath period of specialization, τ.

Maximum length of specialization in the no-loans case was approximately 6.3 years and occurred for the parameter set $\beta_0 = 50$, $\beta_1 = 0.2$, $\beta_2 = 0.3$, $\delta = 0.06$, $r = 0.07$, $R/P = 1.25$, $E_0 = 2{,}000$, and $t_n = 50$. Maximum τ also was obtained for this parameter set and was about 2.7 years. If we assume the age of decision, t_0, to be 18, these numbers would correspond to job market entry at about age 21 for the Ben-Porath individual and a realization of income over and above schooling expenses at about age 24 for the individual with no access to loans for educational investment. Minimums were found for the set $\beta_0 = 30$, $\beta_1 = 0.2$, $\beta_2 = 0.3$, $\delta = 0.07$, $r = 0.09$, $R/P = 0.75$, $E_0 = 4{,}000$, and $t_n = 40$. The minimums were, respectively, $t^* = -0.25$ and $\tau = -2.0$.[20]

Table 3 gives potential earnings (Y_t^*) and net earnings (Y_t) streams for maximum and minimum t^*'s and τ's and their corresponding parameter sets, assuming $t_0 = 18$.

At least for the parameter sets represented in table 3, peak earnings were rather insensitive to loans versus no-loans educational investment. In constrast, length of "full time schooling" is quite sensitive to the loan market assumption. Thus, the major difference in earnings streams for two individuals having equal opportunity other than access to loans during "formal schooling" lies in the early part of work life. Of course, discounted value of earnings at the time of decision favors the individual having access to educational loans, but earnings tend to converge later.

---

[20] The individual cannot, of course, undo the past. The negative figures indicate e.g., that the individual would have entered the labor force at, say, age 16 in the Ben-Porath case with an expected work life of 42 years and a smaller initial human capital stock.

TABLE 3

POTENTIAL AND NET EARNINGS ($1,000):
PATHS CORRESPONDING TO $t^* = 6.3$ AND -0.25, AND $\tau = 2.7$ AND -2.0

AGE ($t_0 = 18$)	$t^* = 6.3$		$\tau = 2.7$		AGE ($t_0 = 18$)	$t^* = -0.25$		$\tau = -2.0$	
	$Y^*(t)$	$Y(t)$	$Y^*(t)$	$Y(t)$		$Y^*(t)$	$Y(t)$	$Y^*(t)$	$Y(t)$
25	24.4	3.0	27.5	6.1	20	5.7	2.0	5.9	2.2
30	42.2	20.9	44.5	23.2	25	9.1	5.4	9.2	5.5
35	55.2	34.3	56.9	36.0	30	11.5	7.8	11.6	7.9
40	64.7	44.2	65.9	45.5	35	13.1	9.5	13.2	9.6
45	71.2	51.8	72.2	52.7	40	14.1	10.8	14.2	10.8
50	75.3	57.6	75.9	58.3	45	14.6	11.7	14.6	11.7
55	76.6	62.3	77.1	62.8	50	14.4	12.4	14.4	12.4
60	74.6	65.6	75.0	66.0	55	13.0	12.5	13.0	12.5
65	67.2	65.0	67.5	64.0	58	11.2	11.2	11.2	11.2
68	58.7	58.7	60.0	60.0					

VI. Summary

For a Ben-Porath-type model of human capital accumulation with the alternative assumption that the individual may not borrow for educational investment, a reduced-form equation was derived that relates the length of "formal schooling" to market, endowment, and ability parameters. The form of the relationship is invariant within the class of homogeneous production processes for augmenting human capital stock except for the manner in which the rental-price ratio, R/P, enters as a determining factor. Thus, the equation may serve as a guide to functional forms and to choosing determining variables in empirical investigation of time spent in school or age of labor force entry.

It was established analytically that the length of "formal schooling" varies directly with the length of work life and inversely with a market rate of discount, r, and the initial stock of human capital. For a Cobb-Douglas production function, it is also generally valid that the rental-price ratio, R/P, affects the length of "formal schooling" in a positive direction; a similar directional effect holds for a Hicks-neutral "ability" index. An experiment involving several specific parameter sets indicated that length of "formal schooling" varies inversely with rate of deterioration of human capital.

A similar reduced form was derived for the original Ben-Porath (1967) set of assumptions. The period of specialization is always shorter for individuals with access to educational loans but who otherwise face identical market, endowment, and technical data. It is generally true that a larger initial endowment of human capital shortens length of specialization in the Ben-Porath case. Other results based on a computational experiment indicate that the individual with access to educational loans specializes longer for larger rent-price ratios (R/P), smaller interest and deterioration rates (r and δ), longer work life, and greater "ability" in the Hicks-neutral sense. Thus, the directional effects of these parameter changes on length of "formal schooling" are invariant for the two models, at least over the range of parameter space considered.

The numerical experiment indicated that length of specialization is sensitive to the loan market assumption. However, earnings streams were mainly affected in early life, with peak earnings (both potential and net) being quite comparable for both assumptions, other things equal.

Appendix

Derivation of the Solution to Ben-Porath's Phase I

In the symbology of the present paper, the objective in the Ben-Porath model was to maximize

$$J = \int_{t_0}^{t_n} \Lambda \, dt, \tag{A.1}$$

where

$$\Lambda = e^{-rt}[R(E_t - K_t) - PD_t] - \lambda_{1t}[\dot{E}_t - \beta_0 K_t^{\beta_1} D_t^{\beta_2} + \delta E_t] \\ - \lambda_{2t}[E_t - K_t - g^2(t)].$$

The Euler-Lagrange conditions are

$$Re^{-rt} - \delta\lambda_{1t} - \lambda_{2t} + \dot{\lambda}_{1t} = 0, \tag{A.2}$$

$$\lambda_{1t}\beta_0\beta_1 K_t^{\beta_1 - 1} D_t^{\beta_2} = Re^{-rt} - \lambda_{2t}, \tag{A.3}$$

$$\lambda_{1t}\beta_0\beta_2 K_t^{\beta_1} D_t^{\beta_2 - 1} = Pe^{-rt}, \tag{A.4}$$

$$\dot{E}_t = \beta_0 K_t^{\beta_1} D_t^{\beta_2} - \delta E_t, \tag{A.5}$$

$$E_t - K_t = g^2(t), \tag{A.6}$$

$$2\lambda_{2t} g(t) = 0. \tag{A.7}$$

Phase I

During the specialization phase, $g(t) = 0$, $\lambda_{2t} \neq 0$, and

$$E_t = K_t. \tag{A.8}$$

From (A.4) and (A.8), it follows that

$$\frac{\dot{\lambda}_{1t}}{\lambda_{1t}} = -r + (1 - \beta_2)\frac{\dot{D}_t}{D_t} - \beta_1 \frac{\dot{E}_t}{E_t}. \tag{A.9}$$

If we substitute equation (A.3) into equation (A.2) to eliminate λ_{2t}, some rearranging results in

$$\frac{\dot{\lambda}_{1t}}{\lambda_{1t}} = \delta - \beta_0\beta_1 E_t^{\beta_1 - 1} D_t^{\beta_2}. \tag{A.10}$$

From equation (A.5), dividing by E and rearranging terms yield

$$\frac{\dot{E}_t}{E_t} = \beta_0 E_t^{\beta_1 - 1} D_t^{\beta_2} - \delta. \tag{A.11}$$

Combining (A.9), (A.10), and (A.11), we get

$$\frac{\dot{D}_t}{D_t} = \frac{\delta + r - \delta\beta_1}{1 - \beta_2}, \qquad t_0 \leq t \leq \tau. \tag{A.12}$$

Thus, during Phase I, the growth rate of the D input is constant.[21] Therefore, the optimal path for the D input during Phase I is

$$D_t = c_1 e^{\gamma t}, \qquad t_0 \leq t \leq \tau, \tag{A.13}$$

where γ is the positive constant on the right-hand side of (A.12) and c_1 is a constant of integration as yet undetermined.

Substituting (A.13) for D_t in (A.11) sets up a differential equation governing the optimal path for E_t (and K_t) during Phase I. The resulting equation is

$$\dot{E}_t - c_1^{\beta_2}\beta_0 E_t^{\beta_1} e^{\gamma\beta_2 t} + \delta E_t = 0, \tag{A.14}$$

which has the solution

$$\log\left[\frac{c_1^{\beta_2}\beta_0\delta(1 - \beta_1)}{\gamma\beta_2 + \delta(1 - \beta_1)} e^{\gamma\beta_2 t} - \delta E_t^{1-\beta_1}\right] = -\delta(1 - \beta_1)t + c_2, \tag{A.15}$$

where c_2 is another constant of integration. An alternative expression for (A.15) is

$$E_t^I = \left[c_1^{\beta_2}\frac{\theta}{\delta} e^{\gamma\beta_2 t} - \frac{c_2^*}{\delta} e^{-\delta(1-\beta_1)t}\right]^{1/(1-\beta_1)}, \qquad t_0 \leq t \leq \tau, \tag{A.16}$$

where

$$\theta = \frac{\beta_0(1 - \beta_1)\delta}{\gamma\beta_2 + \delta(1 - \beta_1)} \quad \text{and} \quad c_2^* = e^{c_2}.$$

One of the integration constants can be eliminated by the assumption that the individual begins the process with an exogenous stock of human capital E_0. By taking initial time to be $t_0 = 0$, (A.16) yields

$$c_2^* = c_1^{\beta_2}\theta - \delta E_0^{1-\beta_1}. \tag{A.17}$$

Phase II

Phase II is identical to both models, with the exception of its point of initiation. We symbolize the end of Ben-Porath's period of specialization by τ.

Determination of τ in the Ben-Porath Model

Equations (A.16) and (A.17) hold at the point τ, and, further, since $y_{2\tau} = 0$,

$$\frac{\beta_1 D_\tau}{\beta_2 K_\tau} = \frac{R}{P}, \tag{A.18}$$

and since τ marks the end of the period of specialization, it is also true that

$$K_\tau = E_\tau. \tag{A.19}$$

---

[21] The growth rate of the D input is constant, but the absolute level of its optimal path will be seen to be a function of input prices as well as other parameters.

Using equations (A.16), (A.17), (A.18), and (A.19), where K_τ is equation (27) evaluated at τ, we get an implicit function relating the length of specialization to the parameters of the model as follows:

$$1 - \delta e^{[\gamma\beta_2 - \delta(1-\beta_1)]\tau} - \frac{\delta}{\theta}\beta_0\beta_1\left[\frac{1 - e^{(r+\delta)(\tau - t_n)}}{r+\delta}\right]$$

$$+ (\beta_0\beta_1)^{-(\beta_2/\Delta)}\left(\frac{\beta_1 P}{\beta_2 R}\right)^{[\beta_2(1-\beta_1)]/\Delta} E_0^{1-\beta_1} e^{-\delta(1-\beta_1)\tau}\left[\frac{1 - e^{(r+\delta)(\tau - t_n)}}{r+\delta}\right]^{-\beta_2/\Delta} = 0. \tag{A.20}$$

References

Becker, Gary S. *Human Capital.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1964.

Ben-Porath, Yoram. "The Production of Human Capital and the Life Cycle of Earnings." *J.P.E.* 75, no. 1 (August 1967): 352–65.

———. "The Production of Human Capital over Time." In *Education, Income, and Human Capital*, edited by W. Lee Hansen. Studies in Income and Wealth, vol. 35. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1970.

Friedman, Milton. *Capitalism and Freedom.* Chicago: Univ. Chicago Press, 1962.

Friedman, Milton, and Kuznets, Simon. *Income from Independent Professional Practice.* New York: Nat. Bur. Econ. Res., 1954.

Haley, William J. "Human Capital Accumulation over the Life Cycle." Ph.D. dissertation, North Carolina State Univ., 1971.

———. "A Generalization of the Path of Human Capital Accumulation." Econometrics Workshop Paper no. 7204. East Lansing: Michigan State Univ., 1972.

Johnson, Thomas. "Returns from Investment in Human Capital." *A.E.R.* 60, no. 4 (September 1970): 546–60.

Lillard, Lee. "An Explicit Solution to the Human Capital Life Cycle of Earnings Model and Its Application to Earnings Distribution." Ph.D. dissertation, North Carolina State Univ., 1972.

Sage, Andrew P. *Optimum Systems Control.* New York: Prentice-Hall, 1968.

Schultz, Theodore W. "Investment in Human Capital." *A.E.R.* 51, no. 1 (March 1961): 1–17.

Some Problems in the Use of Income-contingent Loans for the Finance of Higher Education

---

Marc Nerlove

Northwestern University

In 1945 Friedman and Kuznets presented a radical proposal for financing professional education which Friedman 10 years later extended to higher education in general. Their proposal was, in essence, that individuals sell "shares" in their future income streams in order to finance investment in themselves. This paper discusses why and how the Friedman proposal must be modified in order to take account of information costs and certain political and social consequences involved in "risk-rating" individuals. The recently introduced Yale Tuition Postponment Option is used to illustrate a number of difficult problems involved in actual implementation of the Friedman proposal. It is shown that the consequences of income-contingent loan repayment programs and their financial viability in the absence of subsidy depend crucially on a number of empirical aspects of the behavior of participants and potential participants which are, at present, poorly understood.

> Who goeth a borrowing goeth a sorrowing. [Thomas Tusser, *Five Hundred Points of Good Husbandry* (1573)]

1. Introduction

In *Income from Independent Professional Practice*, Friedman and Kuznets (1945, p. 90) presented a radical proposal for the finance of professional education, which Friedman (1955) later extended to higher education in general. Because the early statement of Friedman and Kuznets put so

---

Much of this discussion appeared as an unpublished appendix (mimeographed, September 1971) to Nerlove (1972). Subsequent to the preparation of my earlier material, a comprehensive analysis of income-contingent loan programs has appeared (Johnstone 1972). Johnstone's book, however, frequently does not make clear the underlying economic structure of the problems involved. This article is directed to the elucidation of this structure. The research on which this paper is based was partially supported by a grant from the Rockefeller Foundation to the University of Chicago for the study of the economics of population and family decision making and partially supported by funds attached to the Cook Professorship in Economics at Northwestern University, which I held on a visiting basis in 1973–74. I am indebted to an anonymous referee and to Robert J. Barro, Gary Becker, William J. Bowen, Stephen Dresch, Stanley Fischer, Robert J. Gordon, T. P. Schultz, T. W. Schultz, and James Tobin for helpful discussions. They are not, of course, responsible for the opinions expressed or for errors in my analysis.

[*Journal of Political Economy*, 1975, vol. 83, no. 1]

© 1975 by The University of Chicago. All rights reserved.

concisely and cogently the essential argument for the income-contingent loan, which is the subject of much recent discussion, it is worth quoting at length:

> Investment in professional training will not necessarily be pushed to the margin because earning power is seldom explicitly treated as an asset to be capitalized and sold to others by the issuance of "stock." An individual ... may borrow money, obligating himself to repay the principal and to pay interest.... Under such conditions, an investor who loaned money to a prospective professional man could at most get back his capital and the interest on it; he could never realize a "capital gain." But he could, and frequently would, suffer a "capital loss".... For this reason, it would be profitable for an investor to finance the professional training of individuals with no resources other than their expected future incomes only at a rate of interest that would be sufficiently high to provide for capital losses as well as for the usual interest charges. Such a rate of interest would probably exceed the expected return from investment in training even though the latter were well above the market rate of interest. On the other hand, if individuals sold "stock" in themselves, i.e., obligated themselves to pay a fixed proportion of future earnings, investors could "diversify" their holdings and balance capital appreciations against capital losses. The purchase of such "stock" would be profitable so long as the expected return on investment in training exceeded the market rate of interest. [Friedman and Kuznets 1945, p. 90].[1]

The essential elements in this first formulation of the concept of income-contingent loans for professional education, later extended to higher education more generally, are: (1) rectification of capital market imperfections in the market for investments in human capital due to the peculiarities of that form of capital and the legal and institutional obstacles which prevent that form from being treated as loan collateral, as physical and financial capital may be; and (2) correction of imperfections in the market for risk bearing which prevent risks on investments in human capital from being reduced through equity financing and pooling in diversified portfolios and/or transferred from borrower and lender to third parties at a price, in the same manner that the risks of fire or loss of life are transferred to insurance companies for an insurance premium. These two elements of the income-contingent loan concept are distinct and separable not only from each other but also from the issue of an element of subsidy to higher education, with which the concept of income contingency is often tied in discussions of the finance of higher education.

---

[1] Actually, the remark that such investments would be profitable as long as the expected return exceeded the market rate is not quite right. It assumes that the "securities" involved, namely, future professional incomes, are individually subject to risks which can be completely eliminated by pooling, or at least reduced to the level of an efficient portfolio utilizing all available capital assets in the economy. Clearly, this may not be possible, since professionals, such as engineers, may be subject to economy-wide risks of fluctuations in future incomes which are correlated across individuals, and therefore cannot be completely eliminated in a "portfolio" which includes only such claims against future incomes.

A current discussion of the income-contingent loan repayment concept often fails to make clear its essential general analytical structure and the way in which different programs combine conceptually distinct elements of that structure. In this article, therefore, I first outline a general analytical framework for the analysis of abstract income-contingent loans for the finance of higher education. Within that framework, I then discuss the Yale Tuition Postponement Option with some reference to the earlier proposal for an Educational Opportunity Bank (Shell et al. 1968; Shell 1970).[2] In conclusion, some rather general problems are noted primarily related to the absence of firm empirical knowledge about the behavior of individuals confronted with various possibilities for the finance and acquisition of the investment in human capital represented by a higher education.

2. General Characteristics of Income-contingent Loan Repayment Plans

If all capital markets were perfect in the sense that all assets were perfectly divisible, that information regarding the uncertainties attached to streams of income from these assets were costless and available to everyone, and that there were no transactions costs or taxes differentially affecting different forms of income, then we know from the famous Modigliani-Miller theorem (see Fama and Miller 1972) that the firm would be indifferent between debt and equity financing. Yet Friedman and Kuznets suggest that there is a considerable difference for the student considered as a "firm" about to undertake a major investment in productive human capital. They argue that, because of the additional risk inherent in human capital as a result of legal restraints on its use as collateral, the risk premium necessary to attract loan funds to such investments would be inordinately high. It seems clear, however, in view of the Modigliani-Miller result, that the promised "dividend payout rate" as a proportion of future income would also have to be correspondingly high, so that the expected returns or costs would be the same in either case for a given degree of risk, from the standpoint of either the lender or the borrower.

The problem of financing an investment in higher education is a deeper one than the issue of debt versus equity. Friedman and Kuznets hint at the central issue when they speak of diversification as a means of balancing capital gains and losses. Indifference between debt and equity financing is really a consequence of perfect capital markets in which information regarding the probability laws governing the future income streams attached to different assets is costless and readily available to all. Taxes on certain forms of income, but not on others, may also play a significant role in the choice between debt and equity and in the detailed formulation of a method of finance, but are perhaps not crucial to central problems of costly information and relatively high transactions costs associated with investments, which may be large to the borrower but which are quite small to a potential lender. A complicating factor of great importance in the analysis of investments in human capital arises from the fact that, unlike business firms, which are more or less out to make a pecuniary profit, individuals may realize a greater part of the benefits from investment in higher education through higher psychic, or nonpecuniary, income.

---

[2] Based on the proposal of the Panel on Educational Innovation (Zacharias 1967).

In capital markets for the equities and debts of business firms, a great deal of information is typically available concerning the income prospects of the larger units, and there are consequently large differences in the rates of return which must be offered at any given point in time in order to attract purchasers for either bonds or shares. Very small firms, or new firms, and especially those entering a new field, suffer serious difficulties in raising capital primarily because the costs of gathering information about their income prospects are prohibitive, not only relative to the amount of capital involved but also often absolutely. The case for the existence of such capital market imperfections is all the clearer for a young student, often unaware of his potential to absorb and utilize the capital represented by a higher education and often undecided on his future occupation or career. Moreover, while risk rating of corporation bonds and stock recommendations and ratings by advisory services are well-established customs in the securities markets, such risk rating and independent appraisals of income prospects for individuals, or even for broadly defined groups of common socioeconomic background or race, would have far-reaching, indeed politically and socially intolerable, consequences, quite apart from the high informational costs involved. Various institutions, such as venture capital funds and investment banking, have arisen in markets for nonhuman capital to reduce the high informational costs through specialization in gathering the requisite knowledge and to pool risks to the degree possible in the absence of the information required for optimal portfolios. While it is possible to view the recently introduced Yale Tuition Postponement Option, or the proposed Educational Opportunity Bank, as a kind of venture capital fund for investments in human capital, it is also possible to view these programs from the standpoint of the borrower as a combination of loan and a form of low-income insurance. Looking at the matter in this way reveals a number of problems related to the financial viability of such programs in the absence of subsidy and in the responses of student borrowers both in choosing whether or not to enter such a program (ex ante) and in choices made once committed to the program (ex post).

Because risk rating of student equities or loans would be enormously costly on an individual basis and fraught with serious political and social consequences on a group basis (even for seemingly innocuous classification criteria, which would be less effective in separating risk classes the more innocuous they are), students are, by and large, classed together in so-called repayment cohorts, each of which is charged common repayment rates and is subject to similar terms and provisions. It is the concept of such a cohort which allows implementation of the idea that, in borrowing under the program, students in effect sell shares in their future income streams without at the same time necessarily tying the scheme to an outside source of subsidy. Membership in a repayment cohort may begin at some time after the student has borrowed the funds to finance a portion of his tuition and other expenses incurred in his education. A provision is invariably made for the borrower to "opt out" of a plan, that is, to terminate his obligation ahead of his repayment cohort, by paying a penalty of some sort for the privilege. The opt-out provision may be expressed in a variety of equivalent ways. For example, in the Yale Tuition Postponement Option, repayment at a rate of interest applied to the balance outstanding of the loan of any cohort member is applied instead to 150 percent of the loan. More simply expressed in the Shell et al. (1968) variant of the Educational Opportunity Bank, it is merely repayment of principal at a higher rate of interest than that charged the cohort generally. Clearly, the Yale variant is equivalent to a very high opt-out rate of interest applied to the actual balance outstanding early in the repayment period, but one which falls over time.[3]

---

[3] For example, let the standard interest rate be 8 percent. Then an individual who chooses to repay his loan in the first year pays the equivalent of 62 percent interest on his loan. In general, assume that the standard rate is r, the opt-out provision expressed as a multiple of the original loan is ϕ, and the proportion of the outstanding balance repaid each period is $p(t)$, which will, in general, vary over time for an income-contingent repayment scheme. Then the equivalent rate of interest, i, for an individual who opts out after T periods may be found by solving the polynomial equation

$$P\left(\frac{1}{1+i}\right) = 1 - \sum_{t=1}^{T} \phi \frac{(1+r)^t}{(1+i)^t} \prod_{j=0}^{t} [1 - p(j)] = 0,$$

where $p(0) = 0$ for the root $[1(1+i)]^*$ closest to 1. It can be shown that all the roots of $P(z)$ have a modulus less than 1, so that the solution $[1(1+i)]^*$ corresponds to the lowest implied rate of interest equating the stream of repayments over time on a loan of ϕL to L, where L is the original loan plus accrued interest at the time of entry into a repayment cohort, provided the smallest root is real. Given the nature of the repayment stream, this can be demonstrated, so that the implied rate of interest is always a positive real number. Moreover, provided $\phi > 1$, it is always greater than r but tends to r as $T \rightarrow \infty$. Some examples of such implied rates of interest implicit in the opt-out provisions of the Yale plan are discussed in Section 4.

In the case of ordinary bonds, there is often a clause preventing the bonds from being called before a specific date or dates and, if repaid, guaranteeing payment above par. Such provisions are designed to protect the bondholders from a decline in the general rate of interest which would make refinancing of the obligation desirable for the borrower. There seems to be no counterpart in equity finance; major shareholders, or others in control of the firm, can always repurchase its shares, thus concentrating their holdings, subject only to the rules of the Securities and Exchange Commission concerning insider trading. Presumably they would wish to do so only if they considered the current price of the security low (its yield high) in relation to future earnings prospects. The opt-out provision characteristic of most proposed income-contingent loan schemes is most nearly analogous to the protection of bondholders through a provision requiring redemption at more than par prior to maturity. Since, however, the Yale Tuition Postponement Option and most subsequent proposals provide for a variable rate of interest (the reasons for which are discussed at some length below), this cannot be the rationale for the opt-out provisions typically included. Nor is prevention of repurchase, except at some penalty, consistent with the notion of equity financing of the acquisition of human capital. The reason is quite different, and to understand why such a provision is universal in income-contingent loan schemes for the finance of higher education, it is useful to look at these plans as a combination of an ordinary loan with graduated protection, or insurance, against low income. The combination of loan plus insurance is designed to compensate, at least in an approximate sense, for the infeasibility of full individual risk rating implied by pure equity financing.

Because students cannot be individually rated, they must be combined into groups for which rating is, in some sense, practicable not only economically (i.e., in relation to the cost of gathering the necessary information) but socially and politically as well. This means that high incomes achieved by some members of the group, and their higher-than-average payments, must be used to offset low incomes earned by others and their consequently lower-than-average payments. In examining the behavior of a prospective borrower and his decision, one can think of him as purchasing low-income insurance at a "premium" which is reflected by the conditions under which he can relieve himself of the obligations imposed by the plan prior to the general contractual termination of his group. The earlier those with relatively high-level income streams opt-out, the greater will be the total repaid, including interest, by the lower-income recipients remaining. It is the need to attract potentially high-income recipients to a repayment cohort without necessarily identifying them individually or offering them separate terms which gives rise to the need for an opt-out provision.

The excess of the opt-out rate of interest over the rate charged those who participate in the plan to the termination date for their repayment cohort is, in a sense, the excess of premium over benefit received paid by those with incomes high enough to make it desirable for them to leave the cohort prior to termination, for the "insurance" they bought in entering the plan against the possibility of such low incomes that repayment of the original loan, plus interest at the general cohort rate, could constitute a severe economic burden.[4] Conversely, those who fail to repay the original loan plus interest charges at the time of cohort termination, because their incomes have been low relative to others in their repayment cohort, are collecting benefits of the insurance they bought, in effect, when they entered a plan with opt-out provisions. In general but somewhat imprecise terms, those with higher-than-average income streams for the cohort as a whole pay at an implied rate of interest greater than those who achieve income streams below average. Early termination and failure to repay principal plus interest at the general cohort rate are extreme cases of ex post payment of premiums or collection of benefits under the provisions of the plan. It should be clear, however, that the income insurance implicit in the plan's provisions for termination of payments, whether by opt-out or cohort termination, is ex ante in character. In this connection, it is important to distinguish between the ex ante problem of adverse selection into the program and the ex post choices of people who do join the program as to the mix of pecuniary and nonpecuniary income they may receive and, more important, their incentives for postschool investments in their stock of human capital. The incentive to shift to nonpecuniary income is what I have called the "moral hazard," by analogy with the usual terminology of the insurance literature (Nerlove 1972, p. S188).[5]

Clearly, opt-out provisions designed to attract potentially high-income recipients and hold them in the program for a sufficiently long period (two conflicting objectives) may be designed in a great variety of ways. The precise manner in which such opt-out provisions are specified will affect ex ante selection to the program and ex post behavior of participants.

---

[4] Indeed, such early termination is automatic under the Yale plan.

[5] In any income-contingent loan repayment plan, repayments of any individual are contingent on his pecuniary income stream over time. What income is, in this connection, varies somewhat from plan to plan and is considerably complicated by the problem of nonworking wives or participants with only partial labor force attachment. This complication, however, is only symptomatic of a deeper and more fundamental difficulty: no matter how one treats the problem of women, occupational choice, or any other aspect of the measurement of income on which repayments are based, that income can only be related to pecuniary income and never to the total pecuniary and psychic satisfactions an individual achieves as the result of his investment of the amount lent him under an income-contingent plan. To the extent that postschool investments in an individual's stock of human capital are important in affecting this distribution, they too may be greatly affected by the incentives set up within an income-contingent loan repayment plan.

The opt-out provision, as noted, may always be expressed as a provision for any individual to repay his loan at a higher rate of interest than his repayment cohort is paying as a group, and this rate, in turn, may be fixed over time for the initial balance (as in the Educational Opportunity Bank) or may vary with the amount of the loan already repaid at the general rate (as in the Yale plan). Alternatively, the tax rate, or rate of repayment as a percentage of income per dollar borrowed, might also be manipulated, in relation to any opt-out provision, in order to attract and hold potentially high-income recipients in the plan. All plans proposed to date provide only for constant tax rates over a fixed or variable cohort repayment period, but there is no inherent reason why this should be the case. The difference between a variable tax rate and a fixed opt-out provision is much like the difference between an ordinary tax and a lump-sum tax. The latter alleviates the incentive of those who do, or could potentially, achieve high pecuniary incomes to shift to nontaxable forms, and also allows those who wish to do so, at some point in time, to take on the full risks of future income variability. Because, however, an appropriately graduated tax rate and a no-penalty opt-out provision could be made to yield the case of an ordinary long-term loan, with no provision for low-income protection, it would be useful in future work to analyze this case as well.

The key features and parameters of an income-contingent loan repayment plan are thus: (1) membership in a repayment cohort; (2) a "tax rate," say τ, on pecuniary income, defined in some fashion, which may differ for different classes of members of the cohort (e.g., men and women); (3) an internal rate of interest, say r, on outstanding balances of members remaining in the repayment cohort, which may (as in the Yale plan) or may not (as in the Educational Opportunity Bank Plan) vary over time (I have more to say below on the problem of variable internal rates of interest); (4) a termination date or repayment period for the cohort, say T, after which time those remaining in the cohort are subject to no further repayment taxes on their incomes; and, finally, (5) a provision for opting out which may always be expressed as a provision for any individual to repay his loan at a higher rate of interest, $R > r$. The repayment period, T, for a cohort as a whole may be fixed or variable; if variable, T is usually subject to a maximum, which may be a very long period, for example, 35–40 years, and may fall below the maximum, as in the Yale plan, only to ensure that the plan does not make a profit at the expense of the members as a group of a given repayment cohort. (There seems to be no reason, however, why transfers from one cohort to another should not take place; they have not, so far, been considered.)

Given an expected income profile for a typical member of a repayment cohort, a repayment period of T years, and a repayment tax rate, τ, per dollar borrowed, it is possible to calculate what such a typical borrower will repay and whether, at any point in time prior to T, it will pay him to opt-out at the higher rate, R. The calculation can be put in terms of a rate of return on the funds advanced by the plan; if this rate of return equals the going interest rate at which the plan can borrow long term for T years plus administrative costs, the plan just breaks even; if it falls short, the student borrowers are being subsidized as a group. Alternatively, one can fix the rate of return to the plan, which may or may not involve some subsidy, and examine the relation among the repayment tax rate, term, and opt-out rate *for a fixed-income profile* and given a particular rule for the treatment of different subgroups of the cohort (e.g., married women). The important point to note is the crucial dependence of the results on the set of income profiles assumed for each cohort. *Moreover, the income profile for the representative borrower may be greatly affected by the terms of the income-contingent loan repayment plan.* This is true both ex ante and ex post: ex ante, the effect constitutes the problem of adverse selection into the plan; ex post, the effect constitutes the "moral hazard" implicit in the whole income-contingency idea that an individual borrower's choice between pecuniary and nonpecuniary income may be affected by the terms of the plan. Unfortunately there is very little empirical evidence on these profiles and effects.[6]

[6] The calculations of Shell et al. (1968) concerning the rates of return that would be realized by the Educational Opportunity Bank for various repayment tax rates, opt-out rates, and terms, given a particular treatment of married women, depend on a particular income profile held fixed in all of the calculations. The construction of these profiles is described in Shell et al. (1968, pp. 29–30). Single and married women are distinguished, and account is taken of different educational attainments within each cohort, so that when the rate-of-return calculations are done, allowance can be made for less borrowing by those who do not finish college and for the particular manner in which the repayment tax on the income of married women is assumed to be assessed. There will, of course, be considerable variation around any typical or representative income profile; those who achieve exceptionally high incomes may be expected to opt out at some point prior to the standard termination date. Shell et al. (1968) handle this problem by computing an entire distribution of income (by deciles) for each level of educational attainment and each age category. (Actually this calculation is only for males; the necessary distribution for women is derived from these calculations.) Incomes in money terms are assumed to grow through time at a steady 4 percent rate. Such growth is presumably the result of both inflation and productivity growth. The income distributions used in the Shell study are not reproduced there, so that one cannot replicate their results without repeating the detailed calculations resulting in the income distributions used. In a subsequent paper, Shell (1970) did attempt to take into account the possible adverse effects of high opt-out rates on the distribution-of-income profiles of participants, but his assumptions seem quite arbitrary. At a difference of 2 percent between the opt-out rate and the going rate, Shell assumes that 100 percent of those in the bottom decile will participate, 95 percent of those in the second decile will, and so on, so that 55 percent of those in the top decile do. Without knowing what the distribution of income assumed actually is, one cannot calculate exactly the reduction in the average income of the cohort so affected. However, a typical income distribution would show a fairly small reduction in the mean income of participants under Shell's assumption because high-income individuals are so rare in the first place. Dresch and Goldberg (1972), on whose work much of Johnstone's (1972) discussion in chapters 4 and 5 is predicated, utilize income classes by age derived from each of three educational attainment categories from a cross-classification of income by

TABLE 1

THREE HYPOTHETICAL "REAL" INCOME PROFILES FOR IVY LEAGUE COLLEGE GRADUATES, UNADJUSTED FOR INFLATION OR ECONOMY-WIDE PRODUCTIVITY GROWTH

Year Following Graduation*	Low Earner ($)	Medium Earner ($)	High Earner ($)
1	7,000	10,000	10,000
2	7,750	10,875	12,500
3	8,500	11,750	15,000
4	9,250	12,625	18,000
5	10,000	13,500	22,000
6	10,600	14,800	25,000
7	11,200	16,100	27,500
8	11,800	17,400	31,000
9	12,400	18,700	35,000
10	13,000	20,000	40,000
11	13,700	21,800	45,000
12	14,400	23,000	50,000
13	15,100	25,400	55,000
14	15,800	27,200	60,000
15	16,500	29,000	65,000
16	17,150	31,500	75,000
17	17,800	34,000	82,500
18	18,450	36,500	90,000
19	19,100	39,000	95,000
20	19,750	41,500	100,000
21	20,400	44,000	100,000
22	21,050	46,500	100,000
23	21,700	49,000	100,000
24	22,350	51,500	100,000
25	23,000	54,000	100,000
26	23,000	54,000	100,000
27	23,000	54,000	100,000
28	23,000	54,000	100,000
29	23,000	54,000	100,000
30	23,000	54,000	100,000
31	23,000	54,000	100,000
32	23,000	54,000	100,000
33	23,000	54,000	100,000
34	23,000	54,000	100,000
35	23,000	54,000	100,000

* Assuming no period of graduate or professional schooling during which payments might be suspended.

3. An Analytical Framework

In table 1, three hypothetical age-income profiles unadjusted for inflation and economy-wide productivity growth are presented. These profiles are

---

education and age derived from the 1970 census and based on 1969 money incomes (p.71). They reduce these data to 15 profiles which are inputs to their computer routine for determining the viability of a plan (i.e., whether or not it can break even for given values of the parameters described in the text). Their analysis does take into account inflation and general economy-wide productivity growth but does not consider how representative the set of profiles they develop may be for the membership in any repayment cohort or how these are affected by the terms of the plan under consideration. Johnstone (1972) also frequently mentions the importance of the income profiles of cohort numbers (e.g pp. 67, 83, 85, 97, 105–6, 132–33, 145), and he discusses in general terms (pp. 110–14) the problem of adverse selection. He suggests that the problem be handled by "risk rating."

perhaps not too fanciful for a student completing 4 years of college at a prestigious institution, such as one of the Ivy League schools. Three profiles are given—one for a low earner, one for a medium earner, and one for a high earner—and these profiles extend over a 35-year period (which happens to be the maximum repayment period envisaged in the Yale plan).[7]

The figures in table 1 represent what various typical graduates would earn at different ages if there were no general rise in labor income over time and no inflation. Table 2 shows the result of superimposing a combined 4 percent rate of inflation and productivity growth on the constant dollar, no productivity growth, and age-income relations.[8] A tax rate that will just amortize a given loan at some rate of interest, r (e.g., equal to the rate at which the plan can borrow or that rate less the subsidy), for a student following the constant-dollar–no-productivity-growth profile will amortize the same loan sooner for a student following any of the other profiles. If the opt-out rate, therefore, is set equal to the rate of return, many students will opt-out before full term. In this case, the plan will lose money since there will be some students following still lower income profiles. If inflation and productivity growth affect all students equally and if the effect of these two factors is large enough, it may happen that nearly every one opts out early. If the opt-out rate is set sufficiently above the interest rate, it will have the effect of holding high-income earners in the repayment program longer than would otherwise be the case, and would thereby permit the plan to break even

---

[7] Generally, the rise in earnings with age and experience is assumed to be due to the interaction between schooling and postschooling investments in human capital (see Mincer 1970, 1972). There may also be some effects due not only to greater investment in human capital, as represented by an investment in an Ivy League college education, but also an important interaction with general ability and/or socioeconomic background. Thus, although, according to the Mincer theory, earnings might be expected to peak later for the medium and high earner than for the low earner, this has been reversed in table 1 (low peaks at 25 years and medium and high at 20 years following graduation). The implicit assumption is that the interaction between innate ability and/or the "right connections" is more important for this type of education and population than for the general population. As will be shown, the present profiles have the advantage of causing an opt-out condition to occur for two out of the three profiles under the Yale plan prior to the statutory termination date (35 years), given plausible assumptions about the rate of inflation and economy-wide productivity growth. Clearly, however, one must be careful not to conclude too much about any plan from such profiles. Not only may they be unrepresentative, but without attaching frequencies to the types of income profiles (these three or a larger set), one cannot calculate the variable repayment period provided in such plans as the Yale Tuition Postponement Option, and thus the actual ex post costs and benefits to different groups.

[8] This may seem rather low in the light of recent inflationary experience, but the 4 percent figure may not be far off in the long run and is consistent with the 8 percent internal rate of interest assumed in the discussion of the Yale plan below. To base the discussion on more recent experience would require an 8 percent rate of inflation and productivity growth and 12 percent, or even higher, internal rates of interest. A simple computer program is available for calculating examples involving such alternatives not presented here.

TABLE 2

THREE HYPOTHETICAL NOMINAL INCOME PROFILES DERIVED FROM THE "REAL" INCOME PROFILES IN TABLE 1 ASSUMING A 4 PERCENT COMBINED RATE OF INFLATION AND GENERAL PRODUCTIVITY GROWTH

Year Following Graduation*	Low Earner ($)	Medium Earner ($)	High Earner ($)
1	7,280.00	10,400.00	10,400.00
2	8,382.40	11,762.40	13,520.00
3	9,561.34	13,217.15	16,872.96
4	10,821.19	14,769.46	21,057.45
5	12,166.53	16,424.81	26,766.36
6	13,412.38	18,726.72	31,632.98
7	14,738.44	21,186.50	36,188.12
8	16,149.11	23,813.10	42,425.64
9	17,649.07	26,615.93	49,815.91
10	19,243.18	29,604.89	59,209.77
11	21,090.52	33,560.10	69,275.43
12	23,054.86	36,823.74	80,051.61
13	25,142.61	42,292.87	91,579.04
14	27,360.49	47,101.60	103,900.59
15	29,715.57	52,227.36	117,061.33
16	32,121.63	58,998.91	140,473.59
17	34,672.63	66,228.62	160,701.79
18	37,376.31	73,942.30	182,323.49
19	40,240.82	82,167.12	200,150.67
20	43,274.68	90,931.61	219,112.31
21	46,486.87	100,265.80	227,876.81
22	49,886.79	110,201.22	236,991.88
23	53,484.33	120,771.06	246,471.55
24	57,289.85	132,010.16	256,330.42
25	61,314.24	143,955.16	266,583.63
26	63,766.81	149,713.37	277,246.98
27	66,317.48	155,701.90	288,336.86
28	68,970.18	161,929.98	299,870.33
29	71,728.98	168,407.18	311,865.15
30	74,598.14	175,143.47	324,339.75
31	77,582.07	182,149.20	337,313.34
32	80,685.35	189,435.17	350,805.87
33	83,912.77	197,012.58	364,838.11
34	87,269.28	204,893.08	379,431.63
35	90,760.05	213,088.81	394,608.90

* Assuming no period of graduate or professional schooling during which payments might be suspended.

despite the presence of some borrowers who never requite their loans plus interest. Inflation and productivity growth affecting all groups uniformly will not affect this conclusion even if the result is to cause many to opt out early; indeed if the opt-out rate is high enough, the plan may do better than break even.[9] But this assumes that those with high-income profiles will, in fact, join the program in the first place and that the repayment tax rates and opt-out rates are not set on the basis of unrealistically optimistic assumptions about the rates of future inflation and productivity growth.

---

[9] Doing better than breaking even is impossible in the Yale plan formulation, since it explicitly provides for cohort termination when the cohort as a whole, including those who opt-out, had repaid the cohort principal plus interest at the plan's internal rate. Indeed, it is worth pointing out here that, if a plan is formulated in such a way as to break even at just a given termination date, or as in the Yale plan some time before then for safety's sake, and it does run to conclusion, it is probable that the lender will have lost money. If the lender loses, the cohort as a whole must gain, and thus the average cohort member obtains a subsidy or a transfer from the lender.

The preceding analysis suggests that there is a positive relationship between the repayment tax rate and the rate of return earned by the plan and a negative relationship between the opt-out rate and the tax rate for a given group of participants with a given distribution of income profiles; that is, for such a group, the repayment tax rate will have to be higher the higher the rate of return required to break even, and lower the higher the opt-out rate. It is difficult to characterize the effect of the distribution of age-income profiles in any simple way, since the interaction between tax rate, opt-out rate, and returns to the plan depends on the whole shape of each profile and the way in which they shift over time as a result of inflation and productivity growth. It is nonetheless useful to simplify the discussion by subsuming all the effects in one measure of central tendency, say the median income at some age. Denote this income by Y and let, as before, τ be the repayment tax rate per dollar borrowed, R the opt-out rate, and r the rate of return earned by the plan. What might be called the break-even relationship among these variables is then defined as that function of τ, R, and Y which gives the rate of return earned by the plan:

$$r = B(\tau, R, Y), \tag{1}$$

where

$$\frac{\partial B}{\partial \tau} > 0, \qquad \frac{\partial B}{\partial R} > 0, \qquad \frac{\partial B}{\partial Y} > 0.$$

With r fixed, (1) defines a relationship among τ, R, and Y such that for a fixed-repayment tax rate, the higher the median income Y, the lower the opt-out rate will have to be in order for the plan to break even at the rate of return r.

The break-even relationship depends, in a complex way, on the distribution of underlying age-income profiles and on the precise treatment of the incomes of different groups in the cohort. Nonetheless, it is possible to give some intuitive arguments about the relationship. No matter how high the opt-out rate, provided it is finite it will not be possible to achieve a given rate of return if the average income of the cohort is zero[10] or even if it is positive, if it is very small. Similarly, if opt-out rates are very low or no different from the going interest rate, average cohort income may have to be very high in order to achieve a given rate of return for a particular repayment tax rate, since those who have high-income profiles in the cohort will quickly opt out, leaving those with low incomes to continue footing the bill.

---

[10] Since income over a working lifetime must be nonnegative, zero average income entails zero dispersion of income and, hence, zero income for each member of the cohort.

Of course, it is clear that *we cannot consider the population of those who enter the program to be independent of the terms offered.* The excess of the opt-out rate R over the going interest rate at which students can borrow to finance their educations is the ex ante "insurance" premium which they pay for the assurance that, if their incomes should prove very low, they will not be burdened with excessive repayments. The appropriate "going rate of interest" to consider in this connection is not the kind of rate we observe on student loans under present imperfect capital market circumstances but, rather, something closer to a conventional mortgage rate which would exist if the market for student loans were made more perfect. As a first approximation, this may be taken to be the rate of return earned by the plan, that is, the rate at which the plan can borrow plus administrative costs, so that no element of subsidy to student borrowers complicates the analysis. The larger the difference between the opt-out rate and the going rate, the higher the cost of the income insurance to those entering the loan program. When the cost is very low, we would expect many to avail themselves of the opportunity; when the cost is high, fewer will feel they can afford the insurance. Who will be those who fail to participate in the program as the cost of income insurance increases? Presumably, those potential participants who feel more certain of above-average incomes, or those who like to gamble, will be the ones who do not join the program when the cost of income insurance is increased. Failure of the former to join will reduce the "average" income of the cohort to which the repayment tax rate and opt-out rate will later be applied. Thus, a demand-for-income-insurance curve may be defined that relates the opt-out rate to cohort income for a given going rate of interest or, equivalently as argued above, to the rate of return earned by the plan.

As with the earlier discussion of the break-even relation, we might express the demand for insurance as

$$Y = D(R, r), \tag{2}$$

where

$$\frac{\partial D}{\partial R} < 0, \qquad \frac{\partial D}{\partial r} > 0.$$

Unfortunately, not much can be deduced a priori about the shape and slope of D in relation to B. In principle, the break-even relationship is calculable for a specified distribution-of-income profiles, but the demand-for-insurance relation depends on the distribution of those who do *not* join the program as well as those who do. Thus, even to say something empirical about the relationship will be difficult. Moreover, the operation of an income-contingent loan repayment program and its viability depend crucially on the demand for insurance, that is, on the extent to which high opt-out rates cause those with the best income prospects not to join the program. It can be shown that what is important for the operation of the plan is the slope of the demand for insurance relative to the slope of the break-even relationship in the vicinity of the point at which the two cross in the $R - Y$ plane.

Substituting (2) into (1), we obtain

$$r = B[\tau, R, D(r, R)]. \tag{3}$$

Holding r constant in (3), we obtain a relationship between τ and R by differentiating (3) totally and solving for $d\tau/dR$:

$$\begin{aligned} \frac{d\tau}{dR} &= -\left(\frac{\partial B}{\partial R} + \frac{\partial B}{\partial Y}\frac{\partial D}{\partial R}\right)\Big/\frac{\partial B}{\partial \tau} \\ &= \left(-\frac{\partial B}{\partial R}\Big/\frac{\partial B}{\partial Y} - \frac{\partial D}{\partial R}\right)\left(\frac{\partial B}{\partial Y}\Big/\frac{\partial B}{\partial \tau}\right). \end{aligned} \tag{4}$$

The first term in parentheses is the slope of the break-even curve in the $R - Y$ plane; the second is the slope of the demand-for-insurance curve. Thus, if the latter is steeper than the former, lower opt-out rates will correspond to higher repayment tax rates, whereas if the reverse obtains, lower opt-out rates induce so many potentially high-income students to borrow that lower repayment tax rates are feasible at the same unsubsidized rate of return. The levels of these two rates affect both the political and economic viability of the income-contingent loan repayment program.

The Zacharias Panel on Educational Innovation (Zacharias 1967) estimated that a repayment tax rate of 0.33 of 1 percent per $1,000 borrowed would be required on 30-year loans and suggested an opt-out rate of 6 percent, but they apparently did not allow for this in their calculations of whether the Educational Opportunity Plan would be self-sustaining at these rates if it could borrow at 4 percent. Shell et al. (1968) assume that the opt-out rate has no effect on the distribution-of-income profiles of participants. Under this assumption and more complicated assumptions concerning the treatment of women's incomes and how incomes might grow, they show that, for opt-out rates of 6.5 percent, tax rates of 0.5 of 1 percent per $1,000 borrowed on a 30-year term, or somewhat less on a 40-year term, would be required to achieve a 5 percent rate of return. At present rates of interest, it is clear that quite high repayment tax rates and/or opt-out rates would be required to achieve a rate of return which would cover both the plan's borrowing and its administrative costs, even on the assumption that the opt-out rate has no effect on the distribution of participants' income profiles. Shell (1970) does make an arbitrary assumption about the effect of the opt-out provisions of Shell et al. (1968) on participation in that version of the plan; under his assumption, there is little effect.

In principle, the matter could be settled for any specific assumption about the effects on each decile of the distribution-of-income profiles, but a realistic appraisal would require that the effects be empirically determined rather than assumed a priori. It would seem then that there are at least two vital gaps in our knowledge bearing on the feasibility and practicality of unsubsidized, or at least not open-endedly subsidized, income-contingent loan repayment plans: (1) the distribution-of-income profiles on which different plans offered by different types of institutions, individually or in concert, can draw, and (2) the effects of a plan's opt-out provisions on the characteristics of those who join. All of this, of course, leaves out the still more conjectural effects of the plan on its members' choices among the various streams of pecuniary and nonpecuniary income and their incentives to invest in themselves in post-schooling forms of human capital.

Since the above factors have been shown to affect repayment tax rates necessary for a plan to break even with no subsidy or a given subsidy, the lack of empirical knowledge is both politically and economically important for two reasons. First, very high repayment tax rates are not likely to be politically feasible. For example, a 1 percent tax rate on income per $1,000 borrowed would cost a family in which both husband and wife borrowed full tuition and board and room costs at an Ivy League college ($20,000 or more per person for 4 years) 30 percent of their income or even more, assuming the wife did not work and was taxed on one-half of her husband's income![11] Such high rates of repayment taxation are probably not feasible and would presumably discourage most people from borrowing so much.[12] Moreover, the rate is added to already large federal, state, and local tax payments.[13] Second, if subsidy elements are introduced into the program, either to make repayment tax rates and opt-out rates more feasible or for reasons entirely extraneous to the program, who participates and who does not will determine who gets the subsidy and who does not. Subsidies for higher education ought not depend on individuals' risk aversion, which is a reason for eliminating them altogether from income-contingent loan repayment programs, if that is possible.

---

[11] This is more generous than the method of calculation suggested in Shell et al (1968), and corresponds to the treatment in the Yale plan discussed below. This program requires that a married woman pay a certain percentage of her own income plus one-half of her spouse's income, if a joint federal tax return is filed. Thus, if no married women worked and all filed joint returns with their husbands, one-half the income of the husband would be subject to the repayment tax. But the possibility that the repayment liability might lead to a shift in the composition of property and labor income and to separate returns as a means for reducing the sum of tax and repayment liabilities is not considered. Provision for opting out does, of course, modify any conclusion one ought to draw about the strength of this effect, as indicated in the discussion earlier in this paper.

[12] I have not discussed the effects of changes in the repayment tax rate on the operation of the program above. Presumably, higher repayment tax rates would work on the distribution of participant's income profiles in a manner opposite that of higher opt-out rates; that is, high repayment tax rates would tend to discourage potentially low-income participants. If any element of subsidy is involved in the program, such discouragement might have undesirable redistributive effects.

4. The Yale Tuition Postponement Option

It is useful to illustrate the foregoing analysis and general description of income-contingent loan repayment programs for the finance of higher education by one which is in actual operation.

The Yale Tuition Postponement Option is a modest income-contingent loan repayment program operated since 1971–72 by a single university for its own undergraduate, graduate, and professional students. It is described in some detail in Johnstone (1972) and is reviewed here only briefly.

The Yale plan was announced in the spring of 1971 by President Brewster in conjunction with a concurrent announcement of major increases in tuition and board and room charges in Yale College and in the graduate and professional schools, both in the current academic year and, prospectively, in future years. The maximum amounts that can be borrowed under the program are closely tied to increases in tuition, board, and room. Currently no traditional forms of aid are available to cover the increase in charges, although in the future this restriction may be relaxed and financial aid continue to be available on the same basis as before to help students meet the charges in effect when the program went into operation.

The Yale plan introduces several novel features into the general framework of income-contingent loan repayment programs developed above. First, the repayment tax rate is set relatively low, 0.4 of 1 percent of adjusted gross income per $1,000 borrowed, with a minimum payment of $29 per $1,000 regardless of income.[14] Second, the term of the loan is variable, with a maximum term of 35 years; a given cohort (students commencing repayment in the same year) may liquidate its obligation sooner if the group as a whole has sufficiently high income. Indeed, present expectations are that obligations will be liquidated in only 26–28 years, the added years being a safety feature designed to protect Yale against loss. Third, the "contract" interest rate which determines both when a cohort's obligation is requited and the terms under which an individual may opt out is variable; the rate is set each 6 months, applicable to the period ending each June 30 and December 31, in such a way as to provide Yale the approximate costs of borrowing plus administration and insurance incurred in operation of the plan. The maximum rate for any 6-month period, however, is limited to the lesser of 12 percent or Yale's cost of borrowing for the coming 6 months plus 1 percent.[15] In general, Yale expects to be able to borrow at close to the prime rate. Until recently this varied between 2.25 and 8.5 percent and was close to 6 percent when the Yale plan was announced; it is now, of course, much higher. Fourth, the opt-out terms set by Yale are stiffer than those envisaged in previous plans and are, moreover, variable; prior to graduation a student may satisfy his obligation by paying the amount he has deferred plus interest compounded at the "contract" rate, plus such insurance premiums as are involved in his contract.[16] Once, however, a student becomes a member of a repayment group or cohort and commences repayment at the fixed percentage of his adjusted gross income, he may not opt out except by repaying 150 percent of his loan plus accumulated interest on that amount compounded at the "contract" rate, which, as indicated, varies over time.[17] If an individual's repayments, as a fraction of his income, ever accumulate to an amount equal to the opt-out figure, he will automatically be excused from further repayments. When a cohort as a whole accumulates repayments equal to 100 percent of the total deferred plus accumulated interest on that amount compounded at the "contract" rate, the cohort as a whole is excused from further repayment, but in no case will the term of obligation extend beyond 35 years. Finally, a married participant who files a joint return may repay on one-half the adjusted gross income reported on that return or on the amount which would have to be reported if the participant were to file a separate return, whichever is greater. Thus, married women who do not work repay on the basis of at least one-half the spouse's income, whereas in such a case the husband, if he were also a participant, would repay on the full adjusted gross income.[18] There are numerous other terms and conditions designed primarily to ensure the collectability of amounts due Yale under the program, but these need not concern us here.

---

[13] Those who argue, such as Grubel and Edwards (1964), that present tax rates have little incentive effect on occupational choices are clearly not talking about evidence related to such nonmarginal changes in the rates at which pecuniary incomes are taxed. The rates implied by the above are far beyond the range of observation of ordinary people, and the behavior of the very rich is surely not too instructive in this connection.

[14] Note that $29 × 35 = $1,015, so that a participant with no or very low income would still, essentially, have to repay the principal of his loan but without interest. The purpose of this provision is to ensure that the amounts paid as interest under the plan are tax deductible. Such deductibility should make the plan very attractive to individuals who may expect very high incomes in later years.

---

[15] The 12 percent ceiling is set by a Connecticut usury law. Presumably, this ceiling (and law) will be changed if interest rates remain long at present levels.

[16] Life insurance covers the amount of outstanding balance plus accumulated interest.

[17] This is in sharp contrast to the Shell et al. (1968) variant in which the opt-out rate is a fixed rate of interest higher than that at which the cohort as a whole repays.

It is useful to illustrate the operation of the Yale plan for three students having the "real" and nominal income profiles given in tables 1 and 2 above. Note that this assumes a 4 percent rate of inflation and productivity growth, which is rather unrealistic at the time of writing. For the sake of illustration, assume that each student borrows exactly $1,000; that he does not begin repayment until 3 years after borrowing (i.e., essentially not until he graduates); that the contract rate remains fixed at 8 percent; and that for this illustration, contrary to Yale's hope and expectation, cohort termination does not take place until a full 35 years have elapsed. Tables 3, 4, and 5 summarize the results for the low earner, the medium earner, and the high earner, respectively. Column 1 of each table shows the year repayment begins (entry to the repayment cohort) to the year in which the cohort terminates or the opt-out condition occurs, whichever is sooner. Column 2 shows the amount repaid in the year in question, 0.4 percent of adjusted gross income subject to a minimum repayment of $29 per $1,000 borrowed. Note that the low earner just barely makes it over the minimum repayment amount in his first year. The next two columns give the amounts from which an early termination date may be calculated. Column 3 is the so-called principal account. It consists of the amount borrowed initially, plus accrued interest at the contract rate (which is 8 percent in this example but normally varies over time in the Yale plan), less repayments which have occurred and are recorded in column 2. The early termination account given in column 4 is calculated in exactly the same way except that the initial amount in year 1, when repayment begins, is taken to be 1.5 times the amount borrowed plus accrued interest. The opt-out condition occurs whenever, in the year following the ordinary amount due, (0.4 percent of adjusted gross income)

---

[18] This may be compared with the proposal in Shell et al. (1968) where the problem of low pecuniary income of women is handled by a minimum-earned-income test. For a married woman earning more than some minimum W, repayment would be based on her income alone; for a woman earning less than W, repayment would be based partly on her income and partly on her husband's. An unmarried woman would presumably be taxed at the same rate as a man, despite the fact that women tend to enter low-paying occupations and tend to receive less than men for comparable jobs. The rate applied to a married woman's income and the fraction of her husband's income taxed would be set in such a way as not to discourage joint filing. The possibility that income from property may be manipulated so as to reduce the joint tax liability is not considered, although Shell et al. specifically suggest that the repayment tax be based on adjusted gross income rather than just income from wages and salaries in order to avoid the general problem of the possibility of shifting between the latter and property income. The problem of the treatment of women is a serious one for any program designed to finance each generation's higher education from its own future income stream.

TABLE 3

OPERATION OF THE YALE PLAN FOR A "LOW EARNER" BORROWING $1,000 3 YEARS PRIOR TO GRADUATION*

Year (1)	Repayment ($) (2)	Principal Account ($) (3)	Early Termination Account ($) (4)
1	29.12	1,331.37	2,011.61
2	33.53	1,404.35	2,139.01
3	38.25	1,478.45	2,271.89
4	43.28	1,553.44	2,410.35
5	48.67	1,629.05	2,554.52
6	53.65	1,705.73	2,705.23
7	58.95	1,783.23	2,862.69
8	64.60	1,861.29	3,027.11
9	70.60	1,939.60	3,198.69
10	76.97	2,017.80	3,377.61
11	84.36	2,094.86	3,563.45
12	92.22	2,170.23	3,756.31
13	100.57	2,243.27	3,956.25
14	109.44	2,313.29	4,163.30
15	118.86	2,379.50	4,377.50
16	128.49	2,441.37	4,599.22
17	138.69	2,497.99	4,828.40
18	149.51	2,548.32	5,065.24
19	160.96	2,591.22	5,309.49
20	173.10	2,625.42	5,561.15
21	185.95	2,649.51	5,820.10
22	199.55	2,661.92	6,086.16
23	213.94	2,660.94	6,359.12
24	229.16	2,644.65	6,638.69
25	245.26	2,610.97	6,924.52
26	255.07	2,564.78	7,223.42
27	265.27	2,504.69	7,536.02
28	275.88	2,429.19	7,863.02
29	286.92	2,336.61	8,205.15
30	298.39	2,225.14	8,563.17
31	310.33	2,092.83	8,937.89
32	322.74	1,937.51	9,330.18
33	335.65	1,756.86	9,740.95
34	349.08	1,548.33	10,171.14
35	363.04	1,309.16	10,621.80

* No opt-out condition occurs. Implied rate of interest, 7.62 percent.

would exceed the amount remaining in the early termination account and, thus, more than requite the balance in that account.

A note on whether the opt-out condition occurs and in what year is given at the bottom of tables 3–5. The low earner pays for the full 35 years in this example. The medium earner opts out in year 29, because in this year his required payment, $673.63, would more than requite the balance of his early termination account, which would amount to $286.24. The high earner opts out even sooner, in year 17, since in this year his payment would have been $642.81, which would have more than paid off the balance of $531.01 remaining in his early termination account. The ex post cost of the low-income insurance or the ex post benefit may be calculated, as suggested in Section 2 above, as the difference between the

TABLE 4

OPERATION OF THE YALE PLAN FOR A "MEDIUM EARNER" BORROWING $1,000 3 YEARS PRIOR TO GRADUATION*

Year (1)	Repayment ($) (2)	Principal Account ($) (3)	Early Termination Account ($) (4)
1	41.60	1,318.89	1,999.13
2	47.05	1,377.35	2,112.01
3	52.87	1,434.67	2,228.11
4	59.08	1,490.37	2,347.28
5	65.70	1,543.90	2,469.36
6	74.91	1,592.50	2,592.00
7	84.75	1,635.15	2,714.62
8	95.25	1,670.71	2,836.53
9	106.46	1,697.91	2,956.99
10	118.42	1,715.32	3,075.13
11	134.24	1,718.31	3,186.90
12	147.29	1,708.48	3,294.56
13	169.17	1,675.98	3,388.95
14	188.41	1,621.65	3,471.66
15	208.91	1,542.48	3,540.49
16	236.00	1,429.88	3,587.73
17	264.91	1,279.36	3,609.83
18	295.77	1,085.94	3,602.85
19	328.67	844.14	3,562.41
20	363.73	547.95	3,483.68
21	401.06	190.72	3,361.31
22	440.80	−234.83	3,189.41
23	483.08	−736.70	2,961.48
24	528.04	−1,323.68	2,670.36
25	575.82	−2,005.39	2,308.16
26	598.85	−2,764.68	1,893.96
27	622.81	−3,608.66	1,422.67
28	647.72	−4,545.07	888.77
29	673.63	−5,582.30	286.24

* *Opt-out condition occurs in year 29 with repayment of $286.24. Implied rate of interest, 10.11 percent.*

rate of interest implicit in the stream of actual repayments and the contract rate of interest, or any other appropriate rate. Of course, such a comparison would be difficult if the contract rate varied, but our example assumes it is constant, so the comparison is straightforward.[19]

---

[19] With a variable contract rate, one would simply compute, from the stream of interest accruals on the balances due of the cohort as a whole, that equivalent constant rate of interest to cohort termination which would just equate the present value of the stream of repayments to the total amount due at commencement of repayment. The complication of this approach arises mainly because it can only be accomplished ex post, after we know who reaches early termination in a plan, and how long the cohort repayment period actually turns out to be. It has been suggested that, instead of using interest rates as the measure of the value or cost to different hypothetical borrowers, the present value of the stream of payments be used. Clearly, the advantage or penalty accruing to a borrower of a deviant rate depends on how long the deviation is maintained. However, the rates calculated in these examples are internal rates of return for the payment streams actually incurred, and are thus precisely equivalent to a present-value calculation and reflect equally the effect of termination equivalent to a stream of zero payments from that date onward. It would thus seem to be a matter of taste whether one compares present values of given streams, computed at, say, the contract rate, or the internal rate of return of the streams.

TABLE 5

OPERATION OF THE YALE PLAN FOR A "HIGH EARNER" BORROWING $1,000 3 YEARS PRIOR TO GRADUATION*

Year (1)	Repayment ($) (2)	Principal Account ($) (3)	Early Termination Account ($) (4)
1	41.60	1,318.89	1,999.13
2	54.08	1,370.32	2,104.98
3	67.49	1,412.45	2,205.89
4	84.23	1,441.22	2,298.13
5	107.07	1,449.45	2,374.92
6	126.53	1,438.88	2,438.38
7	144.75	1,409.23	2,488.70
8	169.70	1,352.27	2,518.09
9	199.26	1,261.19	2,520.27
10	236.84	1,125.24	2,485.06
11	277.10	938.16	2,406.76
12	320.21	693.01	2,279.09
13	366.32	382.13	2,095.10
14	415.60	−2.90	1,847.11
15	468.25	−471.38	1,526.63
16	561.89	−1,070.98	1,086.87
17	642.81	−1,799.47	531.01

* Opt-out condition occurs in year 17 with repayment of $531.01. Implied rate of interest, 10.49 percent.

The low earner receives benefits ex post because he actually repays at a rate 0.38 percentage points lower than the contract rate of 8 percent. Both the medium earners and the high earners pay insurance premiums ex post, since the implicit rate paid is 2.11 and 2.49 percentage points higher than the contract rate.

This example brings out something rather interesting about the Yale plan which can be shown to be generally true of income-contingent repayment plans that have fixed percentage-of-income tax rates and similar opt-out terms. Benefits to low earners, provided they are not absolutely destitute, may turn out to be rather slight compared with a straight loan at the contract rate, whereas the premiums levied on high and medium earners are large relative to the contract rate. The size of the difference, however, diminishes rapidly for the very high earners as compared with the merely well-to-do. Note that opting out at 29 years, only 7 years short of cohort termination, involves an implicit interest rate only 0.37 percentage points lower than opting out 18 years prior to cohort termination. This conclusion would, of course, be modified if, as Yale expects, cohort termination occurs much earlier than the maximum 35-year term. For this reason, and perhaps because of the considerable tax advantages for those who do manage to achieve very high income levels in later years, the Yale plan has proved very popular.[20]

Perhaps the most interesting feature of the Yale plan is the variable rate of interest at which obligations of individuals and cohorts, termination dates, and opt-out terms are computed. The rationale for such a variable rate of interest is that interest rates ought to vary with the rate of inflation and productivity growth in the economy as a whole, a theory whose best-known proponent was, appropriately, Irving Fisher (1867–1947), a Yale man from first to last (Fisher 1954, esp. pp. 399–451).[21]

---

[20] In 1972, 36 percent of those undergraduates eligible took part (*New York Times*, March 4, 1972, p. 56).

Comparison of tables 1 and 2 illustrates "typical" income profiles both in constant dollars and after allowing for steady inflation and productivity growth. One could imagine the constant-dollar profile as shifting upward from year to year, so that an individual could read off his actual income from year to year not from a continuous curve but, rather, from a succession of curves. In reality, of course, neither inflation nor productivity growth will proceed smoothly; hence, if Yale were to base the terms of its tuition postponement option on smooth income profiles which allow only for constant rates of inflation and productivity growth, it is possible that Yale would not break even on the plan in some periods or that students commencing repayment during certain periods might be more heavily taxed than those commencing repayment at different times.

It may be shown that what really matters for the self-sustaining nature of Yale's program, or lack of it, is neither the interest rate nor the rate at which the money incomes of participants grow at given ages but, rather, the approximate difference between those two rates. Let $Y_a(t)$ be the income of a participant of age a in terms of dollars of year t; let τ be the tax rate if the student borrows $1,000; let γ be the rate of growth of money incomes. The participant's income in 10 years in dollars at that time is $Y_a(t + 10)$, where a is the individual's current age (he is assumed to have borrowed at age $a - 10$). If the rate of growth of money income through inflation and productivity growth has been constant at γ,

$$Y_a(t + 10) = Y_a(t)(1 + \gamma)^{10}. \tag{5}$$

Thus, the actual income profile can be constructed from the constant-dollar–no-productivity-growth profile from a series of relationships such as (5). If an amount v of the loan is to be discharged at a rate of interest r (Yale's cost plus administrative and insurance costs), the present value of the payment in year $t + 10$ must be equal to the amount of the loan made 10 years previously which is discharged:

$$\begin{aligned} v &= \frac{\tau Y_a(t + 10)}{(1 + r)^{10}} = \tau Y_a(t)\left(\frac{1 + \gamma}{1 + r}\right)^{10} \\ &\cong \frac{\tau Y_a(t)}{(1 + r - \gamma)^{10}}, \end{aligned} \tag{6}$$

since $(1 + r)/(1 + \gamma)$ is approximately $1 + r - \gamma$ for r and γ both small.

---

[21] Tobin has advanced and elaborated this argument in great detail in an unpublished memorandum, "Interest Rates, Income Growth, and Deferred Tuition" (no date).

Of course, neither the rate of interest nor the rate of growth of money incomes will be constant from period to period, but if their difference is constant, and if Yale bases its break-even repayment tax rate τ on the typical constant-dollar-no-productivity-growth income profile of participants in the program, Yale will, in fact, break even. If the "contract" interest rate were fixed for a given cohort, the repayment tax rate would have to be set higher in order to ensure that the program is self-sustaining; if incomes grew faster than expected in some periods, the entire cohort would be terminated according to the rules, earlier than the full 35-year term, so Yale would not benefit. If incomes grew more slowly, however, Yale might not recoup its investment.

The question, of course, really boils down to whether or not the difference between interest rates and rates of growth of money incomes is approximately constant. This has not been the case, although it has varied in the past a great deal less than money interest rates. What will happen in the future depends on a number of factors, including the extent to which inflation is anticipated and the kinds and effectiveness of policies which the government follows in order to stabilize the economy.

The problem of adverse selection of participants may be greatly compounded by the unilateral nature of the Yale program. It is true that Yale offers a loan program, and the student may make the choice of whether to borrow or not, so that, superficially, it would appear that the student now has a choice he did not have before and must, therefore, be no worse off. But at the present time Yale has also raised tuition and other charges and, for the time being, has denied financial aid on the increment. Thus, a student with good but uncertain income prospects may prefer to attend another institution with lower tuition and other charges, despite the absence of an income-contingent loan repayment program at that institution. The combination of higher current costs and what may amount to a high "income-insurance premium" may have a great effect on the distribution of participants' income profiles, on which the evaluation of the plan and the determination of its parameters so crucially depend. Only further analysis of the Yale experience, however, can reveal how serious a problem this may be.

5. Conclusions

The original suggestion of Friedman and Kuznets for the equity financing of investments in professional education is the origin of recent proposals for, and actual operating programs of, income-contingent loan repayment plans for the finance of higher education. Because the world in which we live, however, is not one of costless information and transactions, nor one in which risk rating of individuals with respect to their potential future income streams is free of social and political consequences, neither equity financing nor long-term loan financing at appropriate rates are practical alternatives. Instead, recent plans and proposals group participants into rather broadly defined repayment cohorts. In order to achieve a measure of transfer of risk offered by equity financing without outside subsidy, the provisions of these plans are designed to hold those who do achieve high incomes in the plan as long as possible without, at the same time, discouraging potentially high-income recipients, who are nonetheless uncertain about their prospects, from participating. It appears useful, therefore, to view such plans as offering long-term loans plus low-income insurance. When this is done, a number of important gaps in our empirical knowledge appear.

First, it is apparent that we know far too little about the behavioral responses to such plans, which combine an insurance element with reduction in capital market imperfections, to assess either their broad social and economic consequences or their financial viability for particular institutions, or groups of institutions acting in concert, with or without governmental assistance. The demand for the "insurance" provided in any income-contingent loan repayment plan depends critically not only on the distribution of future income profiles of potential members of the plan but also on the effect of the terms of the plan on the distribution of those potential members who join and those who do not join.

Second, the effects of the manner in which insurance is received through an income-contingent repayment plan on incentives for achieving income in pecuniary or nonpecuniary form, and on incentives for postschooling investments in human capital, are far from marginal, as some imply, and are certainly difficult to assess quantitatively. This aspect is closely related to the problem of the treatment of married women's incomes and the effects of the terms of the plans on marriage patterns, fertility decisions, and labor force participation.[22]

Finally, it would seem possible to separate the insurance element inherent in an income-contingent loan repayment plan from the aspect of the plan which removes imperfections in the capital market preventing individuals from long-term borrowing to finance investments in their skills and education. Repayment profiles following normal savings patterns and extending over long periods of time, but *not* contingent on income received, would certainly seem feasible under a system of federal guarantees. Reaping the benefits of a higher education is perhaps less risky than farming; yet in 1916 the Federal Farm Loan Act established a series of Federal Land Banks, under the supervision of the Federal Farm Loan Board, which substantially improved the capital market for farmers and enabled them to borrow long term (30–40 years) without external subsidy.[23] Such a plan would not, however, mitigate the economic and social consequences of imperfect (or no) risk rating of individual loans.

---

[22] Shell et al. (1968) attempt a very elaborate formulation, but in the end find the calculations too difficult and base their analysis on a set of rather simple assumptions. The Yale plan incorporates a simple proposal, and it would seem important to understand the effects of similar terms on marriage patterns, fertility, and decisions on female labor force participation, since these would not only affect the operation of the plan itself but would have enormous social consequences were the plan or a similar one to become universal.

Existing income-contingent loan repayment programs are all experimental. We have much to learn from their continued operation, but there is also a great deal which we may begin to infer indirectly through further research on individuals' behavior with respect to investment in various forms of human capital, the risks involved, and the incentives provided by our current system of taxation. In the meantime, other ways of financing higher education without necessary recourse to subsidy deserve to be explored.

References

Dresch, S. P., and Goldberg, R. D. "Variable Term Loans for Higher Education—Analytics and Empirics." *Ann. Econ. and Soc. Measurement* 1 (January 1972): 59–92.

Fama, E. F., and Miller, M. H. *The Theory of Finance.* New York: Holt, Rinehart & Winston, 1972.

Fisher, Irving. *The Theory of Interest.* Repr. 1930 ed. New York: Kelly-Millman, 1954.

Friedman, M. "The Role of Government in Public Education." In *Economics and the Public Interest,* edited by R. A. Solo. New Brunswick, N.J.: Rutgers Univ. Press, 1955.

Friedman, M., and Kuznets, S. *Income from Individual Professional Practice.* New York: Nat. Bur. Econ. Res., 1945.

Grubel, H. G., and Edwards, D. R. "Personal Income Taxation and the Choice of Professions." *Q.J.E.* 78, no. 1 (February 1964): 158–63.

Johnstone, D. Bruce, with the assistance of Dresch, Stephen P. *New Patterns for College Lending: Income-contingent Loans.* Ford Foundation Report. New York: Columbia Univ. Press, 1972.

Mincer, Jacob. "The Distribution of Labor Incomes: A Survey with Special Reference to the Human Capital Approach." *J. Econ. Literature* 8 (March 1970): 1–26.

———. *Schooling, Experience, and Earnings.* New York: Nat. Bur. Econ. Res., 1974.

Nerlove, Marc. "On Tuition and the Cost of Higher Education: Prolegomena to a Conceptual Framework." *J.P.E.* 80, no. 3, suppl. (May/June 1972): S178–S218.

---

[23] The act, its associated institutions, and its consequences are fully described in Taylor (1925, chap. 15). While the act did reduce the interest rates that many farmers paid on land mortgages, some serious economic distortions developed. For example, the real risk of farming in the Texas uplands and western Oklahoma is substantially higher than the risks of farming in Iowa. When federal funds became available at the same rate to both areas, Texas and Oklahoma land prices rose relative to those in Iowa. The analogy in the case of student loans is that high-risk students will be encouraged to undertake heavier investments in higher education than they would if loans were available only at rates that more correctly reflected the risks involved.

Shell, Karl. "Notes on the Educational Opportunity Bank." *Nat. Tax J.* 23 (June 1970): 214–20.

Shell, Karl; Fisher, F. M.; Foley, D. K.; Friedlaender, Ann F.; et al. "The Educational Opportunity Bank: An Economic Analysis of Contingent Repayment Loan Programs for Higher Education." *Nat. Tax J.* 21 (March 1968): 2–45.

Taylor, Henry C. *Outlines of Agricultural Economics.* New York: Macmillan, 1925.

Zacharias, J., chairman, Panel on Educational Innovation. *Educational Opportunity Bank: A Report of the Panel on Educational Innovation to the U.S. Commissioner of Education, the Director of the National Science Foundation, and the Special Assistant to the President for Science and Technology.* Washington: Government Printing Office, 1967.

Communications

The Quantity Theory in the Early Monetary System of West Africa with Particular Emphasis on Nigeria, 1850–1895

Okonkwo A. Nwani

Tuskegee Institute

A number of contemporary writers such as Ida Greaves, Lord Chalmers, E. K. Hawkins, Newlyn, and Rowan have written articles on monetary problems in West Africa. Such articles have concentrated on monetary matters dating from the period of European control of the area. While such writings contribute to knowledge, they leave out an important aspect of monetary problems which is directly related to the cowrie currency system and the "Great Inflation" in West Africa (Nigeria), 1850–95. By so doing, they ignore the famous Marshallian statement: "Natura non facit saltum." The purpose of this paper is to explore the operation of the quantity theory in its rudimentary form in the cowrie[1] currency system of West Africa, and the Great Inflation of 1850–95 in the same geographical and economic region.

The currency of any country constitutes its money. Money is anything which performs the functions of a medium of exchange and a standard of value. In order for anything to fulfill these functions, it must be generally acceptable to sellers or other creditors in settling debts. Money is anything which is generally acceptable and commonly used as a medium of payment at a particular place.

The geographical territory now known as Nigeria had a cowrie currency system and was part of the cowrie currency system of West

Dr. Okonkwo Nwani teaches in the Economics Department at Tuskegee Institute and is from Nigeria. He wishes to thank Jean Federick for typing the manuscript.

[1] Cowries are seashells which were used as commodity money. The unit of currency was defined in "cowrie." They served other purposes such as decorative beads, for offerings during religious worship, etc. Observe that there was no banking system in this area at the time in question.

[*Journal of Political Economy*, 1975, vol. 83, no. 1]
© 1975 by The University of Chicago. All rights reserved.

Africa, which dates as far back as the eighth century. Recent studies recognize its full operation in Nigeria as early as the twelfth century A.D.

Economists writing on money in Africa have left out some important aspects of monetary problems, perhaps because they regarded the cowrie currency as "primitive" money. A Dutch writer in the eighteenth century described money as follows: "What we call money being arbitrary, and its nature and value depending on tacit convention betwixt men, these (Cowrie) shells, in several parts of Asia and Africa, are accounted current money, with a value assigned to them. This is established by a reciprocal consent, and those who are pleased to show a contempt of them don't reflect that shells are as fit for a common standard of pecuniary value as either gold or silver."[2]

The Nigerian cowrie currency, a part of the West African cowrie currency system, was a sophisticated form of currency capable of adjustment to the particular needs of Nigerian trade at the time in question. In West Africa, there are as many words for cowrie (singular) as there are people who used it as money. A few examples will illustrate the point. In Nigeria, among the Ibos, it is "Egoayaka," among the Yorubas it is "Oweyo," and among the Hausas it is "kurdi" or "wuri." In Ghana, among the Ashanti, it is "Cedi." The value of the cowrie was relatively equal in all these areas. The European and Arabic names for this shell were rarely adopted. The cowrie as a currency, and considered with respect to the historical period in which it was used, met the following properties of money, which depended on customs and usage: (*a*) The material was portable, unique, and difficult to counterfeit. (*b*) It was transferable and was readily acceptable as a medium of exchange. (*c*) It was a store of value and a means for deferred payment. With regard to points *a* and *b*, the cowrie as currency was scarce in West Africa and in particular in Nigeria. In monetary history one finds that scarcity was an essential quality of commodity money. To augment the supply of cowries in West Africa, imports became necessary. Two types that were imported were *Cypraea moneta* from the Maldive Islands and *Cypraea annalus* from Zanzibar and the neighboring islands in East Africa. This scarcity made the cowrie currency relatively stable for many centuries before the European influence in West Africa brought instability and the currency's final collapse. The cowrie shells also had a nonmonetary use in this area, as ornaments. It might be interesting to draw a parallel between cowries, and the American Indians' wampum. Wampum served as a medium of exchange and was also worn as an ornament. Both cowries and wampum were strung on threads for various purposes. Both were also used for ceremonial purposes. This is a clear historical coincidence.

---

[2] *A Voyage to the Island of Ceylon by a Dutch Gentleman* (London, 1754), quoted in Gray and Bell (1887, 1890).

The local species of cowries in Nigeria and especially those found along the West African coast were very different from the kinds used as currency and also did not occur in sufficient quantity to serve as money. On the other hand, one could argue that such a currency would be unworkable if money could be picked up in sufficient quantity by residents of the area with every trip to the seashore. The supply of cowries from Zanzibar augmented the available money supply. The nature of cowries made counterfeiting impossible.

On the question of transferability, the cowrie was easily transported in small transactions. Its defect was its relative nondurability, since the cowrie shell was brittle and susceptible to breakage. In addition, many man-hours were needed to count large sums, and the shells, being bulky, were awkward for international trade. The cowrie currency was accepted and is still in use as a regular market currency for small purchases in very rural areas.

A monetary history of cowries is of much significance. As early as the twelfth century A.D., the Nigerian cowrie currency was used widely in the country, which indicates that the economy was well monetized.[3] One finds in Pereira (1936, p. 145): "In the country of Beny [Benin, Midwest State, Nigeria] . . . they use money shells which they call Iguou, a little longer than these 'zimbos' of Maniconguo. They use them to buy everything and he that has most of them is the richest." Thus was the cowrie currency in 1515 described by a contemporary observer.

Barth, in his expedition to Northern Nigeria in 1858, found cowries (*Cypraea moneta*) to be the common currency of the Kano market. In terms of wide circulation and general acceptability, only the cowrie currency had the functional qualities of money, among other commodities such as beads, gin, guns, iron, and strips of cloth, used in small localities.

The wide circulation of the cowrie currency in the West African countries at this early period clearly indicates that the West African economies were monetized. For such an extensive area, there was no need to transport money. The cowrie currency was a unit of account, a common measure of value. Thus it could serve as the unit in terms of which the value of all goods and services was measured and expressed. The cowrie currency had enabled the economy to overcome the inherent problems of the barter exchange system. This theory is in keeping with monetary history, since money became necessary as economies developed.

In many markets literate merchants did not always handle actual cowries but kept detailed accounts in terms of cowries. One such case was reported by Daumas (1848, p. 241) from Northern Nigeria: "You understand that you did not load up with these cowries [in which prices had just been quoted], which were actually only a representative value of the goods sold, of which each of us kept an exact record, where the names of the buyers were written, with their purchase and their debt. . . . . Later we were reimbursed by the exchange of objects equivalent to those which we had brought." This quotation illustrates the function of the cowrie currency as a unit of account. While barter might take place, while other commodities might be exchanged, the prices of such commodities were established in cowrie currency.

---

[3] The importation of cowries as well as their sources are only secondary issues. The primary issue is the use of cowries as currency.

The cowrie currency allowed the operation of discounts for quantity purchases similar to the contemporary discount mechanism. The discount was allowed by the methods of counting utilized. It was almost a decimal system to base 10. The following illustrates how the discount was feasible: 8 × 10 = 100, 8 × 100 = 1,000, 8 × 1,000 = 10,000, 8 × 10,000 = 100,000. Thus 100 was in reality only 80, 1,000 was only 800, 10,000 was only 8,000, 100,000 was only 80,000. The value thus recorded permitted quantity purchase and allowed a profit when retailed. It also allowed a return on lending just as a commercial bank does today with discounts.[4]

Many writers have implied that the reckoning of 80 as 100 is a form of discount designed to favor petty trade or retailing. Einzig (1949, p. 191), whose theory is based on that of Baillaud (1902, p. 70), holds this view explicitly. Profit was possible with relative price stability.

The Collapse of the Cowrie Currency

The cowrie currency had met the needs of the Nigerian economy and had become relatively stable by the eighteenth century, when three major events occurred: (1) large imports of cowries by some European countries into West Africa from 1600 and continuing to 1900, (2) suppression of the use of the cowrie currency by the European colonial governments; and (3) economic inflation in terms of the cowrie currency.

Large Importation of Cowries into West Africa

The scarcity of the supply of cowrie shells in West Africa created a diabolic opportunity for England, Holland, Portugal, and France to trade in cowries. Between 1600 and 1900 these countries bought cowries from the Orient and from Europe. They flooded the West African countries with them. Dutch imports into West Africa are shown in table 1. For *Cypraea moneta* shells the standard figure was 400 cowrie shells to the pound (Troy weight).

During the 1680s, when it had an absolute monopoly of trade, the Royal African Company was exporting well over 1,000 cwt of cowries a

---

[4] See Mage (1868, p. 171). Mage was a French officer who traveled near Segu in 1866. He noted that cowries had a special system of enumeration. The counting was in tens and at first sight was a decimal system.

TABLE 1

IMPORTS OF COWRIES INTO WEST AFRICA, 1689–1780

YEAR	QUANTITY	
	Pounds	Nos. of Cowries
1689	192,951	77,180,400
1721	317,613	127,045,200
1780	133,399	53,359,600

SOURCE.—Martens (1872).

year (Davies 1957, appendix 1, p. 157). In 1717 French merchants estimated their imports to West Africa at 300,000 *livres* a year (Berbain 1942, p. 82). The total export of cowries to Europe and thence to West Africa may therefore have been in the order of many millions a year for the years around 1720. At the beginning of the nineteenth century, the English were shipping some 100 tons of cowries to Africa. By the 1830s, between 60 and 100 tons went annually to Ghana (colonially called the Gold Coast); from 1835, at least 50 tons a year went to the east coast of the Volta (Select Committee on the West Coast of Africa 1842, appendix 36).

Since these areas were contiguous, since they were in the cowrie currency zone, and since trade was carried between them, one may infer how easy it was for an increase in supply in one territory to affect another territory. During the period in question, trade in palm produce, ivory, and other export goods was increasing annually. At the same period the slave trade was at its ignoble peak. The dislocation of the economy due to such social instability initiated and instigated by the Europeans, the decimation of the population, and the excess supply of cowries caused the depreciation of the currency. This was evident by the abnormal rise in prices. The slave trade affected productivity, and, as the quantity theory of money suggests, the increase in money supply without a proportionate increase in output resulted in a rapid rise in the general level of prices. Many historians have engaged in unwarranted disagreements about how cowries reached West Africa. This is obviously not the issue. The vital issue is the inflation that resulted. By the end of the nineteenth century, cowries had fallen in value in parts of the coastal regions to about one-tenth of their value at the beginning of the century.

The Cowrie Currency Zone, 1600–1890

The cowrie currency zone covered the region of the African continent commonly known as West Africa. It extended from Mauritania in the west to Cameroon in the east, and was bounded on the north by the

TABLE 2

NUMBER OF COWRIES TO 1 OUNCE OF GOLD

	Niger Bend	Coast
16th century	8,000	16,000
1700–1800	20,000	32,000
1780–1800	20,000	25,000
1800–1816	20,000	32,000
1816–20	20,000	32,000
1850–60	25,000	38,000

SOURCE.—"The Cowrie Currency of West Africa" (1970).

Sahara Desert and on the south by the Atlantic Ocean. This is the region in which early African empires such as Ghana, Mali, and Songhai flourished.

In contemporary times the area has been politically divided into the following countries: Mauritania, Senegal, Gambia, Portuguese Guinea, Guinea, Sierra Leone, Liberia, Ivory Coast, Mali, Volta, Ghana, Togo, Dahomey, Niger, Nigeria, Chad, Central African Republic, Rio Muni, and Gabon.

The Great Inflation, 1850–95

The exchange rate between the cowrie currency and gold was relatively stable in the area in the early eighteenth century. On the coast the exchange rate was 16,000 cowries to the ounce of fine gold prior to the period of the great inflation (see table 2). Between 1700 and 1820 the cowrie maintained a relatively stable exchange rate with an ounce of gold in the Niger Bend. On the coast the exchange rate fluctuated. Some reasons for the fluctuation were the increases in the importation of cowries and the immediate impact of the slave trade on the coastal region of West Africa, which includes Nigeria. Conditions became worse between 1850 and 1895, the worst period of inflationary pressures. The exchange rate between the cowrie currency and the shilling (sterling), as illustrated in table 3, points out the inflationary spiral.

The great importation of cowries into the area by the European powers led to the rapid fall in the value of cowries (exchange rate) relative to sterling, thus making the value of cowries fluctuate wildly. By 1860 cowries became for a time wholly unacceptable, but later recovered to the value of 1*s.* 6*d.* for 2,000 in Nigeria. After 1863 the value settled down at 1*s.* 3*d.*, falling by 1876 to 1*s.* per 2,000. By 1895, some 2,000 cowries were worth only 6*d.* While one may argue that the exchange rate between cowries and silver was not a clear indication of inflation, commodity prices may serve the purpose.

TABLE 3

VALUE OF 2,000 COWRIES, REDUCED TO SHILLINGS AND PENCE (STERLING), 1850–95: ACCRA, WHYDAH, AND LAGOS

Year	Accra	Whydah	Lagos
1850	4*s*. 2*d*.	3*s*. 8*d*. to 4*s*. 2*d*.	4*s*.
1853	2*s*. 9*d*. to 3*s*.	...	4*s*.
1855	3*s*. 4*d*.	...	2*s*. 6*d*.
1857	...	1*s*. 6*d*.	...
1859	2*s*. 6*d*.	...	1*s*. 6*d*. to 1*s*. 9½*d*.
1863	...	1*s*. 4½*d*. to 1*s*. 9½*d*.	1*s*. 4*d*.
1870	2*s*. 3*d*.	1*s*. 3½*d*.	1*s*. 3*d*.
1874	1*s*. 3*d*.	...	...
1876	1*s*.	...	1*s*.
1881	1*s*.	...	9¾*d*. to 1*s*. 1*d*.
1882	1*s*.	...	6½*d*.
1889	...	6*d*.	...
1895	1*s*.	4½*d*. to 9*d*.	6*d*.

SOURCE.—Burton (1862, p. 143).

TABLE 4

COMMODITIES IN COWRIES, 1850–92

Year	Ox	Kola Nuts	Palm Oil	Hens
1850–74	9,000	2,000	600	200
1876	10,000	12,000	1,200	240
1877	12,000	14,000	1,300	280
1882	15,000	20,000	1,500	300
1884	20,000	30,000	1,600	300
1887	29,000	31,000	1,800	320
1888	30,000	32,000	2,000	330
1889	45,000	35,000	2,500	330
1892	60,000	50,000	3,000	333

SOURCE.—Calculated from data in Burton (1892) and Newbury (1961, p. 42).

Cowries and Commodities

The choice of commodities that may throw light on the inflation problem presents a unique problem of index numbers. An arbitrary choice with some attempt to include a broad range of goods settles down to commercial products that have either domestic or international demand, or both. Palm oil had a high demand in both markets, as did kola nuts, while chicken (hens) had only domestic demand at the period under consideration. Oxen fell under the category of domestic demand. The problem of inflation is illustrated, in table 4, in the prices of these selected commodities expressed in their cowrie currency values during the relevant years.

Table 4 shows that prices for selected products rose rapidly between 1850 and 1892. There was a 566⅔ percent increase in the price of an ox, a 240 percent rise in that of kola nuts, a 400 percent rise in the price of palm oil, and a 66⅔ percent rise in the price of a hen.

The pattern of the price rise for selected products is as follows: for oxen, from 1874 to 1876, prices rose by 11.1 percent; 1876–77, by 20 percent; 1877–82, a 5-year period, by 25 percent; 1882–84, by 33.3 percent; 1884–87, a 3-year period, by 45 percent; 1887–88, by 3.45 percent; 1888–89, by 50 percent; and 1889–92, by 33.3 percent.

For kola nuts, the pattern of the price rise is as follows: 1874–76, 500 percent; 1876–77, 16.7 percent; 1877–81, 42.6 percent; 1882–84, 50 percent; 1884–87, 3.3 percent; 1887–88, 3.23 percent; 1888–89, 9.37 percent; and 1889–92, 42.9 percent.

For palm oil, the pattern is as follows: between 1874 and 1876, prices rose 100 percent; 1876–77, 8.3 percent; 1877–82, 42.6 percent; 1882–84, 6.7 percent; 1884–87, 12.5 percent; 1887–88, 11.1 percent; 1888–89, 25 percent; and 1889–92, 20 percent.

Finally, the price of hens rose as follows: from 1874 to 76, 30 percent; 1876–77, 16.7 percent; 1877–82, 7.14 percent; 1882–84, no change; 1884–87, 6.7 percent; 1887–88, 3.13 percent; 1888–89, no change; and 1889–92, 0.91 percent.

Summary

This paper has shown that the quantity theory of money in its rudimentary form was applicable to the early West African monetary system (cowrie currency). The increase in money supply (cowries) without the proportionate increase in output resulted in price inflation. Also, the collapse of the cowrie currency system was attributable to the large importation of cowries by some European countries into West Africa from 1600 and continuing to 1900, the suppression of its use as currency in the area by the European colonial governments, and the inflation that resulted, leading to a flight from it.

References

Baillaud, E. *Sur les routes du Soudan*. Toulouse, 1902.
Berbain, S. *Le comptoir français de Juda en XVIIIe siècle*. Paris, 1942.
Burton, R. F. *Mission to Gelele*. London, 1862.
"The Cowrie Currency of West Africa." *J. African Hist.* 11, no. 3 (1970): 334.
Daumas, E. *Le grand désert: Itinéraire d'une caravane de Sahara au pays des nègres*. Paris, 1848.
Davies, K. G. *The Royal African Company*. London, 1957.
Einzig, P. *Primitive Money*. London, 1949.
Gray, A., and Bell, A. C. P. *The Voyage of Francis Pyrad of Laval to the East Indies, the Moluccas, and Brazil*. 2 vols. London, 1887, 1890.
Mage, E. *Voyage dans le Soudan occidental, 1863–66*. Paris, 1868.

Martens, E. V. "On the Uses of Shells." *Zeitschrift für Ethnologie* (Berlin, 1872), p. 67.

Newbury, C. W. *The Western Slave Coast and Its Rulers*. Oxford, 1961.

Pereira, Pachecho. *Esmeraldo de situ orbis*. Translated and edited by G. T. Kimble. Brussels, 1937.

Select Committee on the West Coast of Africa. *Report*. London, 1842.

The Equilibrium Size of a Budget-maximizing Bureau: A Note on Niskanen's Theory of Bureaucracy

Albert Breton and Ronald Wintrobe[1]

University of Toronto

I

William Niskanen, elaborating on a hypothesis initially formulated by Parkinson (1962) and Tullock (1965), has recently developed a formal and elegant model of bureaucratic supply of public output (1967, 1971) which implies that a central and characteristic feature of the public sector is oversupply of output and not inefficiency in the narrow sense of output provided at more than minimum cost.[2] Niskanen's model rests on two central assumptions. One is that bureaucrats maximize the size of their budgets and the other that bureaucrats are in effect simple monopolists who are able to impose their own preferences on the governing political party. In this note we would like to raise questions about the two assumptions and suggest that neither of them is acceptable. For that purpose, in the next section, we first restate what we take to be the core of the model in order to bring out some implications which Niskanen overlooks. Then we abandon the assumption that bureaucrats are simple monopolists and replace it with one that allows a role to politicians, but we retain the hypothesis of budget maximization. We show that the model so modified can account for some features of the real world which the pure Niskanen model is unable to explain. In the third section, we go further and suggest that a number of other important facts appear to be inconsistent with the assumption of budget maximization itself. We conclude by arguing that an alternative hypothesis is needed which would be more general, in the sense that it would be capable of providing an explanation for the facts that appear consistent at least with the modified version of the Niskanen model adumbrated here, but which could also explain those facts which seem to be beyond the capacity of even the modified model.

---

[1] Wintrobe's contribution to this note is based on his doctoral dissertation to be submitted to the University of Toronto.

[2] The latter type has been called *X*-inefficiency by Leibenstein (1966).

[*Journal of Political Economy*, 1975, vol. 83, no. 1]
© 1975 by The University of Chicago. All rights reserved.

II

Niskanen begins by characterizing the relationship between a bureau and its sponsor as one of bilateral monopoly, since he assumes that a public bureau is a monopolistic supplier of a service facing a single buyer—the governing political party. But Niskanen goes on to assert that "although the nominal relation of a bureau and its sponsor is that of a bilateral monopoly, the relative incentives and available information, under most conditions, give the bureau the overwhelmingly dominant monopoly power" (1971, p. 30).

The central motivational hypothesis of Niskanen's model is that bureaucrats maximize the size of the budget under their control. This is justified on two main grounds: first, on utility-maximizing ground, since it is asserted that income, prestige, power, emoluments, and other amenities are a positive monotonic function of budget size; and second on survival ground, since pressure from subordinates for larger budgets on the one hand and from executive and legislative committees on the other[3] is said to ensure the survival of budget-maximizing bureaucrats only in the same way that competition among firms dictates the survival of profit maximizers only.

The sponsor's preferences are represented by the marginal valuation curve labeled *CE* in figure 1. Since the preferences of governing politicians are assumed to accurately reflect the preferences of all citizens, *CE* can be taken to portray correctly the sum of the citizens' marginal valuations of the services of the bureau. It follows from the assumption that bureaus are monopolists that they will obtain the budget they prefer, subject only to the constraint that the total costs of supplying the particular service must not exceed the maximum available budget.[4]

The equilibrium rate of output is therefore at the point where the marginal valuation of output is zero—that is, at *E* in figure 1—and the equilibrium budget is equal to the entire area under the "demand" curve, unless the minimum cost of that output exceeds the maximum budget (*OCE*), in which case the bureau is said to be in the "budget-constrained" region. In this region, the equilibrium rate of output is at the point where total cost (the area under the marginal cost, or *AM*, curve) and total budget are equal. This latter case is shown in figure 1: equilibrium output is q_B with a budget of $OAMq_B$ ($= OCDq_B$).

---

[3] It is claimed that in the political system of the United States, legislative committees at least wish and expect bureau managers to request larger-than-required budgets so that they will themselves be capable of performing their own roles.

[4] The constraint implies that the output of public bureaus is always measurable; otherwise a bureau could obtain a larger budget by promising more than it intends to deliver, and the constraint would not be binding. In some cases (e.g., Defense, External Affairs), it is difficult to define the output, let alone measure it, as Niskanen recognizes (1971, pp. 20–22). Breton, Breton, and Wintrobe (1972) deal with this problem at length.

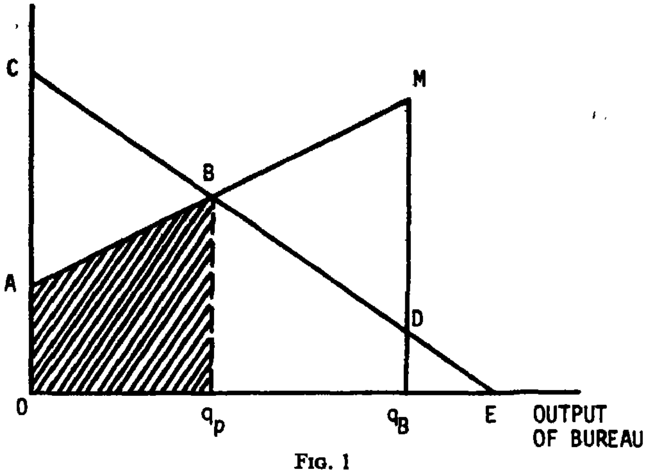

FIG. 1

It follows that a bureau can supply an output which is up to two times as large as would be the rate of output of a competitive firm or industry facing the same demand and cost conditions.[5] Since all bureaus have to be assumed to behave in the same way and to be endowed with the same monopoly power, the size of the public sector and hence the amount of taxes paid by citizens will be up to two times the social optimum, and, consequently, most—if not all—citizens will be in a state of disequilibrium. Now, if there is an idea that is central to economics, especially after the work of Coase (1960), it is that individuals will react to a disequilibrium

---

[5] The socially optimal rate of output in fig. 1 is q_P; it is also the rate desired by politicians under the assumption in the text. As stated, a bureau in the budget-constrained region produces twice this output. It is perhaps easier to see this conclusion algebraically. Following Niskanen's approach, let the bureau's total cost function be

$$TC = cq + dq^2. \tag{1}$$

The sponsor's preferences are represented by a budget-output function, which gives the maximum budget the sponsor will grant for a specified output.

$$B = aq - bq^2. \tag{2}$$

The constraint specifies that $B \geq TC$. The bureau's optimum in the budget-constrained region is where $B = TC$. Solving for q from the equality of (1) and (2) gives

$$q_B = \frac{a - c}{b + d}.$$

The sponsor's optimum is where the marginal benefit from output equals its marginal cost. Differentiating (1) and (2),

$$MC = c + 2dq, \tag{3}$$

$$MV = a - 2bq. \tag{4}$$

Solving for q_P from the equality of (3) and (4),

$$q_P = \frac{1}{2}\frac{a - c}{b + d},$$

which is $\frac{1}{2}q_B$.

situation, whatever its origin, as long as the marginal benefit exceeds the marginal cost of doing so.

We would then have to conclude that at the occasion of the first election the governing political party would be defeated, since the electorate would seek to reduce both the quantity of public goods it receives and the amount of taxes it pays. However, in the nature of the case, the newly elected political party has to be assumed to be unable to exercise its power vis-à-vis the bureaucracy and therefore be unable to implement the promises to cut taxes and public expenditures that have led it to be elected in the first place. That party would also consequently be defeated at the next election. That process would continue without end, one party replacing the other at each election. This does not seem to be a very interesting theorem and does not appear to be supported by the facts.

Building on an idea of Tullock, Oliver Williamson (1967) has recently shown that an inverse relationship between the size of a profit-maximizing organization and the degree of control exercised by its managers over its employees will set a limit to the optimum size of a firm. If, as Niskanen suggests, a bureau's employees can "indirectly influence a bureaucrat's tenure . . . through the real and perceived performance of the bureau" (1971, p. 40), then clearly the advantages to the bureau's manager of larger size will similarly be offset by increasing loss of control over the bureau, and there will be a point beyond which an income- or prestige-maximizing bureaucrat would not seek further budget increases.

The basic flaw in the assumption that bureaus are simple monopolists is really in the counterpart assumption that sponsors are completely passive. Politicians interested in their reelection will obviously attempt to exploit their position as monopsony buyers of the bureau's output. This would seem to leave us in the well-known indeterminacy of bilateral monopoly, but this need be the case only if the problem is formulated to make it that way. Let us examine the kind of results that we get by *holding to the hypothesis of budget maximization, but dropping the* simple monopoly one.

Consider instead that: (1) bureaucratic managers are subject to the exercise of direct hierarchical controls—rewards and sanctions—by the governing political party; a bureau may be a monopoly, but individual bureaucrats are not monopolists, and the careers of senior bureaucrats at least are tied to the sponsor's evaluation of their performance; (2) the limited "power" of a public bureau over its sponsor does not stem from its position as a monopoly supplier of a service, but rather from its control of information. To see this, imagine that information about the minimum costs of a bureau's services is costlessly available to its sponsor. Then if the sponsor constrained the bureau to supply the optimal output at minimum cost, the bureau would have to do so; a bureau head who did not would be replaced, so that the preferences of bureaucrats would be irrelevant in the determination of public output.

The bilateral monopoly framework therefore can easily obscure the fact that the bargaining power of a bureau depends on its ability to distort or conceal information from the sponsor. The problem for governing politicians in controlling this type of behavior is the same as that of the head of any organization in which the interests of subordinates differ from those of superiors, and governing politicians—like other managers—have a battery of techniques at their disposal to help them in that task. Some of these have been described and examined by Downs (1967), including direct monitoring, overlapping bureaus, duplication of services, and the purchase and acquisition of information from alternative sources, including sources at lower levels in the bureau itself.

The general problem of ensuring that subordinates act in the interests of superiors is referred to in the literature on organization theory as the problem of control. Tullock (1965) has invented the term "control-loss" to refer to the cumulative discrepancy between the actions or results of the actions of subordinates and the desires of superiors. In the following discussion, this control framework will be exploited to provide an alternative theory of the equilibrium size of a budget-maximizing bureau.

To that effect, assume that the governing party attempts to maximize the probability of its reelection. In order to isolate the role of bureaucratic preferences in the analysis, assume further that the rules governing the election of politicians are "optimal," in the particular sense that the victorious political party is always motivated to supply public goods to the point where the sum of the marginal benefits to all citizens is exactly equal to the marginal cost of supplying the good. (See n. 10 below, where this assumption is modified.) At that point, total consumers' surplus is maximized.[6]

In figure 1, politicians therefore desire an output of q_P supplied at minimum cost, implying a budget of $OABq_P$. The bureau, on the other hand, will attempt to supply an output of q_B to have a budget of $OCDq_B$. The difference between the budget that the bureau succeeds in obtaining and that desired by politicians is the amount of consumers' surplus appropriated by the bureau and is therefore a measure of control-loss.

Let us now assume that antidistortion or "control" devices are available to politicians at a cost. The cost of the incremental use of these devices is portrayed in figure 2 as the *LR* curve. Controls have many uses, including that of serving to control costs and as a deterrent against distortion of information. Their value to politicians in all uses, however, is simply the reduction in the excess budget of bureaus which their use makes possible.

---

[6] That point is unique only if we possess a social welfare function or some other device that will determine the distribution of income—it is hoped a "just one"—for us.

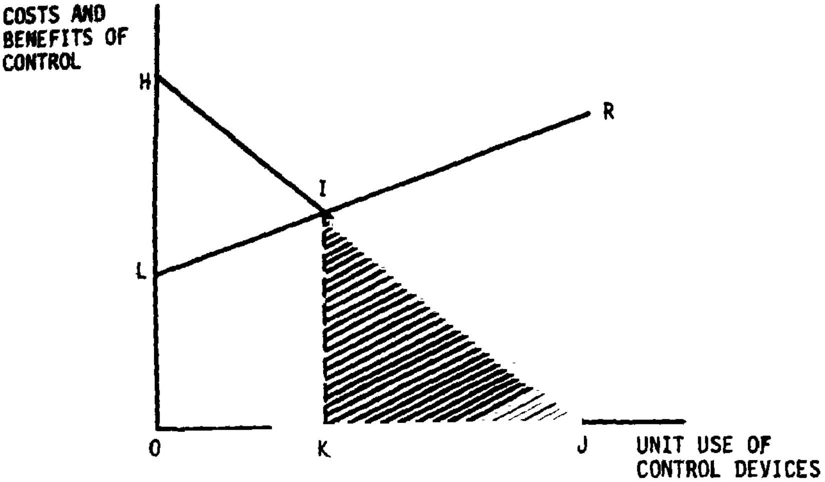

FIG. 2

In figure 2, *JH* is a marginal "revenue" curve which shows the benefits to politicians (and hence to citizens) of a reduction in a bureau's budget brought about by operating control machinery at different levels. The area *OHJ* in that figure is equal in monetary terms to area *ACB* in figure 1, that is, it is equal to the total loss in consumers' surplus if control processes are unavailable.[7]

The sponsor will incur control expenditures up to the point at which marginal benefits—measured along the *HJ* curve—are equal to marginal cost, that is, up to point *K* in figure 2. In equilibrium, total control costs are equal to *OLIK*, and the reduction in the budget of the bureau, using as a starting point the budget desired by the bureaucrats, is *OHIK*. The equilibrium size of the bureau's budget, given the availability and use of antidistortion devices, is therefore equal to the money value of $OABq_P$ in figure 1 plus the money value given by *KIJ* in figure 2—that is, to the sponsor's optimum budget plus the excess which it does not pay him to eliminate. Note that it is the amount of excess budget given by the area *KIJ*, and not literally the area *KIJ*, which is added to the sponsor's desired budget $OABq_P$. Changing the units in which control processes are measured would change the literal size of *KIJ* relative to $OABq_P$, but it would not of course change the sum of money represented by *KIJ*.

---

[7] In fig. 1, the bureau is in the "budget-constrained" region, and hence the loss is due entirely to excess production and not to excess costs of production. Whatever the source of the loss, it must be equal to *ABC* under Niskanen's assumptions. As we make clear later on, in our framework control procedures are assumed to be effective against both sources of inefficiency, but not necessarily equally so.

Alternatively, the equilibrium budget is equal to $OCDq_B$ in figure 1, minus $OHIK$ in figure 2—that is, to the bureaucrats' optimum budget minus the excess which it pays the sponsor to eliminate.[8]

There are two limiting cases which deserve special interest. If control devices are so costly that it does not pay to use them (in fig. 2, *LR* lies entirely above *HJ*), then our results coincide with Niskanen's, and the bureau appropriates the entire consumers' surplus for itself.[9] On the other hand, one could rationalize the traditional assumption about the neutrality of public servants as an assumption that control devices are costless and therefore the excess budget is always zero. Of course, the traditional assumption is not one about the costs of control, but about the motivation of bureaucrats; it is interesting, however, to note the equivalence of costless controls and selfless bureaucrats![10]

The equilibrium output is not depicted in figure 1. To be able to represent that level of output, we have to clear up an additional problem. The extent to which output will exceed q_P will depend on whether the bureau prefers to take the excess budget in the form of larger output or in the form of higher costs on the inframarginal units of output. The extent to which bureaus will adopt one or the other alternative will depend on the relative effectiveness of control procedures in reducing two sources of inefficiency: excess supply and use of inefficient production techniques.

---

[8] It might appear that the optimum output should take control costs into account so that sponsors would simply equate the marginal benefits of output to the marginal costs of output plus those of control. This is not the case, however, since control costs are not a function of output, but of consumers' surplus that would be lost if the costs were not incurred. Hence, they are like fixed costs in that they cannot be imputed to specific units of output.

[9] Where the cost of controls is high, the sponsor may be expected to look for substitutes. The importance of "loyalty" in bureaucracies may be explained in this fashion; where formal control devices are costly, it may be more efficient to hire even less competent but loyal bureaucrats whose careers are tied to the sponsor's rather than to the bureau, and who may therefore be trusted not to distort information in order to obtain a bigger budget. Of course, similar considerations apply not only between sponsors and senior bureaucrats, but between any two levels of the hierarchy where subordinates have some monopoly power over information.

[10] If we remove the assumption that the system of political representation is optimal, the demand curve in fig. 1 (*CE*) may be taken to represent the preferences of governing politicians, but not the sum of citizen's demands for the bureau's output. The analysis in figs. 1 and 2 is unchanged, but q_P is no longer the optimum output from the point of view of the citizenry. For example, if, as Niskanen contends, review committees in the U.S. Congress are biased toward oversupply and the Congress simply follows their recommendations, *CE* would lie to the right of the sum of citizens' demands, and there would be oversupply of output at q_P. On the other hand, if citizens suffered from "tax illusion" or for other reasons vote for a smaller quantity of public output than they in fact prefer, their true demand would be understated by the political process and q_P would be too small. In both cases, of course, the inefficiency is due to the structure of political representation and cannot be attributed to the bureau, a point which Niskanen's formulation obscures. A nonoptimal system of political representation may allow governing politicians some freedom to pursue their own objectives at the public's expense, but there is no reason why they would cede this freedom to the bureau. Hence, the analysis of figs. 1 and 2 always holds.

To see this, assume initially that the cost of control devices is prohibitive and that the bureau is initially budget constrained so that the entire consumers' surplus is appropriated through oversupply of output. Assume now a technological change that makes costless the use of the antidistortion devices that pertain to the flow of information, but does not change the prohibitive cost of supervising and hence of altering the use of inefficient production techniques, so that sponsors can now—as a result of the technological change—monitor actual costs but are still unable to monitor the relation between these costs and the true minimum cost function.

To put it still differently, the sponsor knows the bureau's actual costs, but he has no idea of how efficient the bureau actually is. In such circumstances the best strategy for the bureau is to decide not to operate efficiently and to pretend that its marginal real-resource cost curve is the sponsor's marginal valuation curve but instead to use expensive production processes with declining marginal costs so that its *actual* cost curve shifts upward until it is identical to the sponsor's marginal valuation curve. In equilibrium, the bureau will produce q_P with a budget of $OCBq_P$ and the size of the budget will have been reduced from $OCDq_B$ to $OCBq_P$, but the loss *in terms of consumers surplus* is unchanged from what it was before the reduction in the budget, and the benefits from using control devices are zero.

A necessary condition for control devices to be useful, therefore, is that they be effective to some extent against both sources of inefficiency: the optimal level at which the two types of control mechanisms should be operated is where their marginal productivities are proportional to their respective marginal costs.

A model which takes account of the sponsor's potential use of control processes has at least one implication which is different from that of the Niskanen model: that implication relates to whether oversupply of output or X-inefficiency is the likely source of inefficiency in bureaucratic supply. To eliminate overproduction, the sponsor needs only to acquire information about the bureau's actual cost curve; to reduce X-inefficiency, he needs that information plus an estimate of the true minimum cost of supplying the service. Since it is costlier to police X-inefficiency than to police oversupply, the implication is that the major source of inefficiency in bureaucratic supply is X-inefficiency and not oversupply of output.

The modified model of bureaucratic supply presented above is capable of accounting for some empirical facts which the pure Niskanen model—which has no role for control devices—cannot explain. Since there is no reason to think that the productivity of control instruments is uniform across bureaus, we could hypothesize that in controlling bureaus performing relatively routine tasks, the sponsor's information requirements would be lower, *ceteris paribus*, than in controlling the bureaus involved in more complex or nonroutine tasks, at least in that the information acquired in a particular period retaines its usefulness longer as a deterrent to distortion in routine-task than in non-routine-task bureaus. We would therefore expect the equilibrium use of control instruments to be greatest in bureaus performing relatively nonroutine tasks and using relatively specialized personnel. This hypothesis is consistent with the evidence presented in numerous sociological studies (Blau 1968; Blau, Heydebrand, and Stauffer 1966; Meyer 1968; Woodward 1962) that the ratio of control personnel to direct labor in both firms and government agencies does vary systematically in this fashion.

As we have shown, the monopoly assumption implies that instruments of control are not used because they are too costly to justify their use. Evidence not only that they are used in a wide class of organizations, but that their employment is not haphazard but varies in a systematic and predictable way, is especially difficult to reconcile with this assumption and in particular is difficult to reconcile with the view that sponsors are completely dominated by the bureaucrats. Indeed, the systematic and rational use of control devices implies that the preferences of sponsors, and possibly those of citizens, are reflected in the supply of public output.

Another piece of evidence, admittedly much more impressionistic, supports the view that control devices are used by sponsors to control bureaus and indicates that the supply of public output is not completely at variance with the preferences of politicians and of the public.

Much of the recent interest in the role of the bureaucracy in political decision making was stimulated by the search for an explanation of U.S. involvement in the Vietnamese war. An influential thesis—propounded by, among others, the historian Arthur Schlesinger, Jr. (1967), and the political scientist Richard Barnet (1972) and which Daniel Ellsberg has termed the "quagmire myth" (1972)—is that U.S. military escalation was inadvertent, largely unintended, and prompted by the exaggerated claims and self-interested distortions of evidence by the military to show that each small additional escalation would guarantee success.

Niskanen does not take a position on this issue, but he does specifically cite the conduct of the U.S. military in Vietnam as an example of the workings of the monopoly assumption (1971, p. 76), and furthermore the thesis is surely broadly consistent with his own.

Internal documents and memoranda released as the Pentagon Papers have largely demonstrated this explanation to be false, as Ellsberg has shown (9) and Schlesinger himself has accepted (1971). Estimates prepared by the CIA, the systems analysis branch of the Defense Department, and the federally funded Institute for Defense Analysis were used by the Johnson administration to contradict the exaggerated claims of the military, and the administration's overall estimates of the likely effectiveness of various escalations were surprisingly accurate. If the military budgets were too large, it would not seem to be because the executive branch could not utilize competition between bureaus, redundant sources of information, and other instruments of control to evaluate effectively the claims of the military. It is presumably because such budgets reflected the preferences of the executive, and possibly those of a segment of the public.

III

Finally, we would like to raise a few objections to the assumption that bureaucrats act in such a way as to maximize the size of their budget. As we have already stated, and as Niskanen himself recognizes, such a maximand can only be used if "salary, perquisites of the office, public reputation, power, patronage, output of bureau . . . are a positive monotonic function of the total budget of the bureau during the bureaucrat's tenure in office" (1971, p. 38). While it seems consistent with casual observation and reasonable to assume that a positive monotonic function between the size of budgets and the salary and other amenities of office exists within a given bureau, the same kind of casual empiricism indicates that such a relationship does not exist between bureaus.

To put it differently, the salary of bureaucrats, as well as the other benefits of office, may be larger in some bureaus with smaller budgets than in others with larger ones. The absence of a positive monotonic relationship across bureaus is not of itself catastrophic for Parkinson-Tullock-Niskanen-type models, but when we take into account the additional facts that mobility among bureaus in governmental organization is large, and that heads of bureaus often improve their salaries and other amenities of office by moving from a position at the head of a relatively large bureau to one at the head of a smaller one, this would seem to indicate that a different objective function is being maximized.

In the Canadian government, for example, the salaries (and also, one observes, the power and prestige) of senior bureaucrats in the Department of Finance and the Treasury Board are higher than those of their counterparts in other, much larger bureaus such as Health and Welfare, and a bureaucrat in Health and Welfare could be promoted to Finance at the same time as the size of the budget under his control is reduced.

Another empirical observation which is implied by the assumption of mobility, but which is inconsistent with the hypothesis of budget maximization, is that bureaucrats sometimes cut budgets as a means to further advancement. For example, the present secretary of defense in the United States, James Schlesinger, made his reputation while head of the CIA in part by substantially reducing the agency's personnel![11]

We suggest, therefore, that careful empirical work would reveal that a positive monotonic relationship between budget size and the flow components of a bureaucrat's wealth across bureaus does not exist. Combined with what appears to be a large amount of mobility among the various bureaus of governmental organization, the implication is that there is a great deal of observable bureaucratic behavior which is inconsistent with the hypothesis of budget maximization.

An additional difficulty with the budget-maximization hypothesis is that it is critically dependent on a much more precise definition of a bureau than what Niskanen or anyone else who has used the hypothesis—including Breton, Breton, and Wintrobe (1972)—has provided to date. The problem is that it is difficult to define a bureau in a way that does not automatically involve something essential about the size of its budget, so much so that if one adopts a particular definition of a given bureau its budget will be large and even growing, while on a different definition it will be small and possibly falling. To be sure, falling budgets are consistent with budget maximization as much as falling profits are consistent with profit maximization, but then it is imperative that the consequence of falling budgets be spelled out as they are in the standard theory of the firm.

However, a still more difficult problem for the hypothesis comes from the fact that preoccupation with analytical models of bureaucracy originate in an age when government revenues (and expenditures) have been growing very rapidly as a result of the existence of progressive income taxation coupled with large price increases and rapid economic growth. These three factors, even in the absence of tax illusion, but in the presence of decision rules to elect representatives that are in effect compound simple-majority rules and thus imply the holding of office with the support of only a subset of the electorate, would be sufficient to explain much of the increase in public revenues—and expenditures—that we have observed over the recent half-century. In other words, we are saying that it is not clear what the behavior of bureaucrats would be and what kind of models we would be churning out if government revenues were not increasing automatically.

Although the existence of transfer payments blurs the picture, a look at the behavior of bureaus at the provincial and local level would be rewarding, because these levels of government have not experienced the same automatic increase in their revenues as have central governments.

---

[11] *Time*, February 20, 1974.

Indeed, the so-called centralization of government expenditures analyzed by Peacock and Wiseman (1967) and also by Bird (1970)—but much more skeptically—may simply be the product of the different elasticities of the tax bases of central and local governments. A good theory of bureaucracy would have to be one that is capable of explaining the behavior of bureaucrats whether at the national, provincial, or local level. We suggest that a comparative study of bureaucrats at these different levels of government would find much that can be imputed to the different elasticities of tax bases and much less to the budget-maximization hypothesis.

IV

In this note, we have formulated a model of bureaucratic behavior based on the assumptions that bureaucrats maximize the size of their budgets and that politicians are able to enforce their own preferences by the use of control devices, and it appears that such a model is capable of accounting for not unimportant aspects of the real world of bureaucracy. On the other hand, we have suggested observations and empirical tests that would seem harmful to the budget-maximization hypothesis. One way out of this dilemma would be to recognize that although the hypothesis of budget maximization is not, as we believe, a useful one to employ in general, it may be applicable in certain contexts. We would expect to observe the kinds of behavior associated with budget maximization only where mobility is relatively restricted, which may be the case for the managers of relatively specialized bureaus such as the military. A more satisfactory alternative to the multiplication of models, each valid only in particular contexts, would be to abandon the assumption of budget maximization entirely and to replace it with the assumption that bureaucrats maximize wealth, as the other agents in economic theory are usually assumed to do. Breton (1974) constructed a theory of bureaucratic supply using this assumption in which the constraint on bureaucratic wealth maximization is provided by the sponsor's use of control instruments, as it is in the present formulation. That theory yields a number of interesting empirical implications, including those developed in Sections II and III of this note. It does not, however, provide us with a theory of the equilibrium size of a bureau's budget, as we have attempted to do here using the stronger assumption that bureaucrats maximize the size of their budgets. It is the natural course of progress in economic theory to derive empirical propositions under successively weaker and less restrictive assumptions about human behavior. We hope that the present analysis, which modifies Niskanen's model, contributes to that progress in the economic approach to bureaucracy and will stimulate the interest of others in the subject.

References

Barnet, R. *Roots of War*. Baltimore: Penguin Books, 1972.

Bird, R. M. *The Growth of Government Spending in Canada*. Toronto: Canadian Tax Found., 1970.

Blau, P. M. "The Hierarachy of Authority in Organizations." *American J. Soc.* 73 (January 1968): 453–67.

Blau, P. M.; Heydebrand, W.; and Stauffer, R. "The Structure of Small Bureaucracies." *American Soc. Rev.* (April 1966).

Breton, A. *The Economic Theory of Representative Government*. Chicago: Aldine, 1974.

Breton, A.; Breton, R.; and Wintrobe, R. "A Theory of Public Bureaus." Mimeographed. 1972.

Coase, R. H. "The Problem of Social Cost." *J. Law and Econ.* (October 1960).

Downs, A. *Inside Bureaucracy*. Boston: Little, Brown, 1967.

Ellsberg, D. *Papers on the War*. New York: Simon & Schuster, 1972.

Leibenstein, H. "Allocative Efficiency vs. X-Efficiency." *A.E.R.* (June 1966).

Meyer, M. W. "Expertness and the Span of Control." *American Soc. Rev.* (December 1968).

Niskanen, W. A., Jr. "The Peculiar Economics of Bureaucracy." *A.E.R.* (May 1967).

———. *Bureaucracy and Representative Government*. Chicago: Aldine-Atherton, 1971.

Parkinson, C. N. *Parkinson's Law and Other Studies in Administration*. Boston: Houghton Mifflin, 1962.

Peacock A. T., and Wiseman, J. *The Growth of Public Expenditure in the United Kingdom*. London: Allen & Unwin, 1967.

Schlesinger, A. M., Jr. *The Bitter Heritage: Vietnam and American Democracy, 1941–1966*. Boston: Houghton Mifflin, 1967.

———. "Eyeless in Indochina." *New York Review of Books* (October 21, 1971).

Tullock, G. *The Politics of Bureaucracy*. Washington: Public Affairs Press, 1965.

Williamson, O. E. "Hierarchical Control and Optimum Firm Size." *J.P.E.* 75, no. 2 (April 1967): 123–38.

Woodward, J. *Industrial Organizations*. London: Oxford Univ. Press, 1962.

Property Rights and the Dynamic Inefficiency of Capitalism: Comment

H. E. Frech III

University of California, Santa Barbara

In a recent article in this *Journal*, Lancaster (1973) presented "a new paradigm for capitalism" which provides a new way of looking at the welfare losses involved in "intergroup rivalry" and leads to "a clearly defined dynamic welfare loss."

Incredibly, the inefficiency of Lancaster's model results from the absence of private property rights in the means of production. Thus the model lacks what seems to be the defining characteristic of capitalism: private ownership of the means of production.

Lancaster's Analysis

In Lancaster's model two groups act in concert in dealing with each other. The workers have complete control over the proportion of current output they consume and the proportion saved in any given period. Any income saved is given to the capitalists, who choose to either consume the income or invest it.[1] There is no investment risk involved. The property-rights structure of the model can be detected easily in Lancaster's discussions of the "workers' dilemma," the "capitalists' dilemma," and the causes of the inefficiency in his model.

The Workers' Dilemma

The workers' dilemma arises from the fact that they may provide resources to the capitalists but have no control over whether the capitalists invest the funds or simply use them to support the latter's own current consumption. If the capitalists invest the funds, the workers may benefit by higher future consumption, but if the capitalists consume the resources, the workers suffer a wealth loss.

I would like to thank Axel Leijonhufvud, Wayne Leeman, Gerald O'Driscoll, M. Bruce Johnson, and a referee for helpful comments and discussions.

[1] I use Lancaster's definitions of workers and capitalists. But note that his "workers" actually supply capital, while his "capitalists" only manage it.

[*Journal of Political Economy*, 1975, vol. 83, no. 1]
© 1975 by The University of Chicago. All rights reserved.

The workers have control only over the share of current income that they consume. Their decision problem is, in Lancaster's words: "*Should they forego present consumption by handing over part of total income to the capitalists? If they do not, they will obtain no higher consumption in the future. If they do, they have no guarantee that the capitalists will actually invest sufficient of this income to bring about the desired level of increase*" (p. 1095).

The Capitalists' Dilemma

The capitalists' problem is closely related to the workers'. The capitalists have full control over investment; however, they will not receive the entire amount of future proceeds (since relative income shares are in the control of the workers).

Again in Lancaster's words, the choice problem for the capitalists, arising from worker control of distribution of future income, is: "*Should they spend now, or accumulate in order to spend more later? If they spend now, they know what they have available. If they accumulate, they may fail to obtain their expected share of the increased output when they come to spend*" (p. 1096).

The Result

Setting up a simple differential game model[2] of investment decisions embodying the above features, Lancaster arrives at the predictable solution that the capitalists do not invest enough. Since this model calls for maximum accumulation for some time interval followed by a switch to maximum consumption for another interval, the specific inefficiency arises because "the optimal switch time is *later* than the capitalist switch time" (p. 1102; italics in original).

The Cause of the Loss

The welfare loss in this model arises from the inability of both workers and capitalists to receive the entire future benefits of current sacrifices. Thus, the workers are not willing to give up current income (even when it would be efficient to do so) since they know that the capitalists will not invest all of it. And the capitalists, even if they received more of current income from the workers, would not invest it since they know they will receive only part of the future proceeds. As Lancaster puts it: "The capitalists' valuation of investment is less than the social valuation because the capitalist will receive only a fraction . . . of the increased income generated, while society will receive the whole of it. The workers' valuation of investment . . . does not count . . . because the workers cannot force the capitalists to accumulate against the latter's own best interests. Nor does the fact that the combined value of investment to workers and capitalists together is in excess of [the cost] affect the program, because the groups make their decisions separately" (p. 1105).

---

[2] While the details of the differential game model may be interesting, they are incidental to the purposes of this note. Further, it is clear that with or without sequential decisions, any model of investment behavior with Lancaster's assumptions about workers' and capitalists' property rights would exhibit similar underinvestment inefficiency (Furubotn and Pejovich 1970, 1972).

Lancaster's Paradigm of Capitalism

Lancaster's analysis has little to do with capitalism as that term is commonly, if imprecisely, used. The general idea is that capitalistic economic systems are characterized by private property and free exchange.

Recent literature on the economics of property rights defines private property to include (*a*) the right to control the resources in question, (*b*) the right to enjoy the income stream resulting from the use of the resources in question, and (*c*) the right to transfer the right to someone else (Alchian and Demsetz 1972, p. 783; Furubotn and Pejovich 1972, p. 1140).

If we examine the rights of the workers in Lancaster's model, we see that they have very ill-defined rights to the wealth of the society. If they choose, they may consume all the income in each period. However, they may not use this income to increase future income by investing it themselves. They must give it to the capitalists and hope that they invest it. Clearly, they do not have private property rights to the fruits of their savings. Obviously, they will undersave.[3]

Turning to the capitalists, it is even clearer that they do not have private property rights. After being given resources by the workers, they can consume all of the resources; but if they invest, part of the proceeds will be taken away from them by the workers. (Remember, the workers choose what proportion of the society's current income they consume in each period.)

Essentially, this system gives each group rights only if resources are consumed immediately. If the workers save, they lose rights to their resources. If the capitalists save and invest, they also lose their rights. In such a system where private property rights are attenuated if income is saved and invested, it is not very novel that there will be underinvestment.[4]

[3] This is similar to the "common pool" problem of nonprivate property rights in such goods as oil, fish in public waters, and water resources, which consistently leads to undersaving (too much current consumption). For an important early statement in this large literature, see Gordon (1954).

[4] For analysis of attenuated property rights in nonprofit and socialist firms, see Furubotn and Pejovich (1970; 1972, pp. 1154–56).

The lack of private property rights prevents market interactions from improving the resource allocation. Let us see how private property rights would avoid the problems encountered here and allow the decentralized decisions of workers and capitalists to mesh.

Market Interactions with Private Property Rights

If the workers had private property rights, regardless of their decision to save or consume their resources, many exchange possibilities would avoid Lancaster's inefficiency. Perhaps most directly, if the workers desired to increase their future income, they could simply purchase ownership shares in a firm, which gives them the right to control the firm and also the right to the future proceeds of the investments. Of course, in some sense the workers become the capitalists if this course of action is taken. But, under a private property system, there are other alternatives which preserve Lancaster's separation of the workers' decision to save from the capitalists' decisions to invest. I will outline two, but there are undoubtedly more possibilities.

The workers could simply rent the investment services of the capitalists for some fee. In this case, the behavior of the capitalist is constrained by the terms of the exchange—he is not allowed to simply consume the money which is entrusted to him. Since the workers control all the wealth in the society in Lancaster's model, they might have to pay the capitalists only a small fee (slightly more than the capitalists could have earned as workers).

In this case the workers have private property rights in the means of production, even though they do not make the actual decisions.

More naturally, the workers could make loans to the capitalists at contractually fixed rates of interest. Thus, the capitalists would have the incentive to invest the money in order to repay the loan. In this familiar case, the workers' valuation of investment would be transmitted to the capitalists through the mechanism of the interest rate. The capitalists have full private property rights in the proceeds of investments (they receive the residual), so they will invest if the value of the investment exceeds the cost.

Summary

The inefficiency shown by Lancaster is a direct result of the lack of coordination and incorrect incentives of decentralized systems having no private property rights in productive assets hence no opportunity for coordination through market relationships. His model illustrates interesting, but no longer novel, properties of economic systems having incomplete or attenuated property rights. Examples would include economies with a great deal of theft, frequent political revolutions, publically owned firms, or unrestrained governments (whether democratic or not), which would lead to unpredictable future taxation.

References

Alchian, A. A., and Demsetz, H. "Production, Information Costs, and Economic Organization." *A.E.R.* 62 (December 1972): 777–95.

Furubotn, E., and Pejovich, S. "Property Rights and the Behavior of the Firm in a Socialist State." *Zeitschrift Nationalokonomie* 30 (1970): 431–54.

———. "Property Rights and Economic Theory: A Survey of Recent Literature." *J. Econ. Literature* 10 (December 1972): 1137–62.

Gordon, H. S. "The Economic Theory of a Common Property Resource: The Fisheries." *J.P.E.* 62, no. 2 (April 1954): 124–42.

Lancaster, K. "The Dynamic Inefficiency of Capitalism." *J.P.E.* 81, no. 5 (September/October 1973): 1092–1109.

The Interest Elasticity of Savings, Income Taxes, and the Permanent-Income Hypothesis

Paul J. Taubman

University of Pennsylvania

In the public finance literature the multiperiod consumption savings model has been used to study the efficiency aspects of income, consumption, and other taxes.[1] The multiperiod model has also been used by Friedman, Ando and Modigliani, and others in theoretical and empirical consumption studies, but without any explicit allowance for the effect of taxes on consumption-saving decisions. As shown below, incorporation of income taxes leads to several changes in Friedman's (1957) analysis and in the interpretation of results from cross section studies. Perhaps our most important conclusion is that if cross section studies are to be used to buttress Friedman's contention that the consumption–permanent-income elasticity is one, then the interest elasticity of savings must be zero. A second important result is that cross section data can be used to estimate the interest elasticity of savings.

The Permanent-Income Hypothesis

In the late 1940s and early 1950s many empirical studies of the consumption function were completed. Those studies based on short-period time series or cross section yielded a consumption–current-income elasticity of less than one, while studies based on time series spanning many years yielded elasticities of one. In an attempt to reconcile these results, Friedman (1957) and Modigliani and Brumberg (1954, pp. 388–436) proposed (or reintroduced) some more complex consumption models.[2] Their major assumptions are: a consumer allocates both present and future income to achieve an optimal lifetime consumption plan, and each consumer's utility function is homothetic about the origin.[3] By

The author wishes to thank Irwin Friend for his comments and suggestions.

[1] See Musgrave (1959).

[2] Duesenberry (1949), with whom we shall not be concerned here, has an alternative explanation.

[3] A homothetic utility function is one in which all expansion paths are straight lines. An expansion path connects the points on the various indifference curves with the same slope.

[*Journal of Political Economy*, 1975, vol. 83, no. 1]

© 1975 by The University of Chicago. All rights reserved.

definition, with such a utility function and with a given set of prices, the *elasticity of consumption* with respect to income is one.

The essence of these two models can be captured in a two-period, zero-estate model in which the fairly general constant-elasticity-of-substitution utility function, which is homothetic, is used.

$$U = (C_0^{-d} + C_1^{-d})^{-1/d}, \qquad -1 \leq d < \infty, \tag{1}$$

where C is consumption, d is the elasticity of substitution which is no less than -1, and the subscripts are the time periods.

Friedman maximizes (1) subject to the two-period budget constraint; we will amend Friedman and follow Musgrave (1959) by having the constraint depend on the tax system.

Suppose that there is a proportional income tax.[4] Then the budget constraint can be written as:

$$[(1 - t)Y_0 - C_0 + W_0)][1 + i(1 - t)] + (1 - t)Y_1 - C_1 = 0, \tag{2}$$

where i is the before-tax interest rate, t is the proportional tax rate, and W_0 is beginning-of-period financial wealth.

Maximization of (1) subject to (2) yields

$$[C_1/C_0]^{1+d} = 1 + i(1 - t). \tag{3}$$

In (3) the ratio of future to present consumption is independent of Y_0, Y_1 and W_0, but will vary directly with i and inversely with t unless d equals -1. Since (3) is independent of future and present disposable income, a tax-rate increase will raise the proportion of current and permanent income that is currently consumed, though the elasticity of C_0 with respect to permanent income will be one at the higher tax rate.

Within a cross section of individuals, Friedman and many others rely on the assumption that i and t are either constant or distributed independently of permanent income.[5] But the assumption can be shown to be wrong. For example, Pechman and Okner (1972) have found that the ratio of federal income tax payments to economic income increases continuously for incomes up to $500,000.

The permanent-income theory can be expressed as

$$\ln C = \beta \ln Y_p(1 - t) - \left(\frac{1}{1 + d}\right) \ln [1 + i(1 - t)],$$

---

[4] For simplicity we assume that interest paid is allowed as a deduction from the tax base.

[5] The same assumptions are made in time-series analysis, and the same objection that tax rates have varied can be made here. There is, however, at least a study by Wright (1969, pp. 275–300) in which after-tax interest rates are included in the consumption function.

where Y_p is permanent income and where β is supposed to be equal to one. As long as $i(1 - t)$ is constant, a regression of C on Y_p will yield an unbiased estimate of β. But with $i(1 - t)$ negatively correlated with $Y_p(1 - t)$ and $1 + d$ nonnegative, omission of the last term will yield an estimate of β biased upward.

As far as I am aware, extensive cross section testing of the permanent-income theory has indicated that the elasticity is higher with respect to permanent than to current income.[6] However, no one has yet reliably estimated the permanent-income elasticity to be greater than one.[7] But to the best of my knowledge, in all these tests, including my own, $i(1 - t)$ has not been held constant. But if d is not equal to -1, Friedman's theory would require such estimated elasticities to be greater than one.[8]

It may well be that d is -1, though there is little evidence by which to judge.[9] But if variations in income tax rates do not affect the consumption-savings decision, then an income tax does not impose an excess burden over and above that imposed by a consumption tax.[10] Thus, the efficiency arguments of Harberger (1964, pp. 25–70) and others for a consumption or value-added tax would lose their force.

Let me close on a more positive note. The tax file used by Pechman and Okner (1972) permits us to calculate the average and marginal federal income tax rates by income level and by states. Such tax-rate information could be combined with cross section consumption or wealth surveys to estimate the interest elasticity of savings—a parameter of key concern in many areas of economics.[11]

References

Duesenberry, J. *Income Saving and the Theory of Consumer Behavior*. Cambridge, Mass.: Harvard Univ. Press, 1949.

Friedman, M. *A Theory of the Consumption Function*. Princeton, N.J.: Princeton Univ. Press, 1957.

Harberger, A. "Taxation, Resource Allocation, and Welfare." In *The Role of Direct and Indirect Taxes in the Federal Revenue System*. Princeton, N.J.: Princeton Univ. Press, 1964.

[6] For an extensive summary, see Mayer (1972).

[7] There are a few instances of such an elasticity when housing is used as a grouping or instrumental variable, but many people have objected that, contrary to grouping assumptions, housing expenditures are correlated with taste.

[8] He would also require the elasticity to vary with changes in tax laws.

[9] See, however, Wright (1969).

[10] A consumption tax affects labor-leisure choices. Indeed, since a smaller income tax rate would yield the same revenue as a consumption tax when saving is positive, an income tax may distort labor-leisure choices less.

[11] In making these estimates, it would be best to go beyond the Friedman framework to incorporate uncertainty. However, as long as variations in the subjective discount rate are treated as a random variable. Samuleson (1969) has shown that the uncertainty model in the isoelastic case is similar to the one used here.

Mayer, T. *Permanent Income, Wealth, and Consumption: A Critique of the Permanent-Income Theory, the Life-Cycle Hypothesis, and Related Theories.* Berkeley: Univ. California Press, 1972.

Modigliani, F., and Brumberg, R. "Utility Analysis and Aggregate Consumption Functions: An Interpretation of Cross Section Data." In *Post Keynesian Economics*, edited by K. Kurihara. Brunswick, N.J.: Rutgers Univ. Press, 1954.

Musgrave, R. *The Theory of Public Finance.* New York: McGraw-Hill, 1959.

Pechman, J., and Okner, B. "Individual Income Tax Erosion by Income Class." In *The Economics of Federal Subsidy Programs*, pt. 1. Washington: Government Printing Office, 1972.

Samuelson, P. "Lifetime Portfolio Selection by Dynamic Stochastic Programming." *Rev. Econ. and Statis.* 50 (August 1969): 239–47.

Wright, C. "Saving and the Rate of Interest." In *The Taxation of Income from Capital*, edited by A. Harberger and M. Baily. Washington: Brookings Inst., 1969.

Miscellany

The Economics of Reducing Faculty Teaching Loads: Comment

In a recent article in this *Journal*, McKenzie (1972) sought to demonstrate with elementary indifference-curve analysis the impact of changes in maximum teaching or research loads on the levels of teaching and research undertaken by a faculty member. This note indicates that a basic constraint has been omitted from McKenzie's model and that its inclusion invalidates his results. The analysis will also be extended to incorporate the influence of income, which was omitted from McKenzie's model, on the allocation of faculty time between teaching and research activities.

McKenzie's model may be summarized as follows. Minimize disutility $U(T, R)$ subject to a constraint $T = H - rR$, where T and R are time spent on teaching and research, H is the maximum teaching load, and r is the number of units of T which can be traded for a unit of R. The first-order conditions for this problem are:

$$U_T - \lambda_1 = 0,$$
$$U_R - \lambda_1 r = 0,$$

where λ_1 is the marginal disutility of teaching. The optimal combination of T and R must satisfy the condition that $U_T/U_R = 1/r$; this is depicted by the tangency point between an indifference curve and the technical trade-off locus in McKenzie's diagrams.

Changes in the maximum teaching load, *ceteris paribus*, are equivalent to a change in the level of H combined with a simultaneous change in r which preserves the condition H/r – constant R. Like the income and substitution effects of traditional price theory, an increase in H tends to raise both T and R, while the associated rise in r tends to reduce T and increase R. The net effects of these two influences on the levels of T and R undertaken depend on the form of individual's utility function. In McKenzie's model the "income effect" of successive increases in the maximum teaching load at first dominates the substitution effect, leading to increases in both T and R; eventually, however, the substitution effect exceeds the income effect, and T falls.

Changes in the individual's leisure time are implicit in McKenzie's model because equilibrium positions involving more T and R necessarily imply less leisure. It will be shown that McKenzie implicitly assumes that the marginal utility of leisure is zero; if this is not the case, his conclusions concerning the individual's equilibrium position, and the effect of changes in maximum teaching or research loads, are generally invalidated.

Even if changes in T and R do not affect income, as McKenzie assumes, every individual is confronted by a constraint which indicates that the sum of T, R and leisure time (L) equals the total amount of time available to the individual. Changes in the combined amount of time spent on teaching and research imply corresponding changes in leisure time. Unless the permitted technical trade-off between T and R in McKenzie's model equals unity, a faculty member could reduce the total amount of time spent on teaching and research compatible with the constraint by increasing one activity and decreasing the other. For example, consider McKenzie's optimal point and assume that the slope of his constraint is less than unity, which implies that the faculty member can save time and still satisfy the constraint by increasing R and reducing T. The time saved enables leisure to be increased; this change tends to increase total utility, and if this were the only consideration, complete specialization would be optimal for a faculty member. However, it is also true that the increase in R and reduction in T along the constraint will increase total disutility from R and T. This is implied by the fact that McKenzie's optimal point minimizes total disutility from $T - R$ combinations on the constraint, and is depicted diagrammatically by a move to indifference curves indicating successively higher levels of total disutility as one moves to the right along the constraint away from the tangency point while implicitly holding leisure constant. The faculty member will continue to move along the constraint, increasing R and reducing T, until the marginal utility of the resulting increase in leisure is just equal to the net increase in disutility resulting from the move along the constraint while holding leisure constant. At this optimal point the MRS between T and R will exceed the slope of McKenzie's constraint and will also exceed the slope of the time constraint which equals unity.

These conclusions can be confirmed by adding the time constraint $T + R + L =$ constant to McKenzie's model, and maximizing $U(L, T, R)$ subject to this constraint and the technical trade-off constraint. The first-order conditions for this problem are as follows:

$$U_L - \lambda 2 = 0,$$

$$U_T - \lambda 1 - \lambda 2 = 0,$$

$$U_R - \lambda_1 r - \lambda_2 = 0,$$

where $\lambda 2$ equals the marginal utility of leisure. From the second and third conditions it is clear that the optimal marginal rate of substitution between T and R equals

$$U_T/U_R = \frac{1 - (1 - r)U_L/U_R}{r}$$

only when $U_L = 0$ or $r = 1$ is this condition equivalent to McKenzie's optimal condition.

If teaching and research yield positive utility to the individual, and if a faculty member is limited to McKenzie's constraint, the preceding analysis and its conclusions are equally applicable; the only difference is that the indifference curves are convex to the origin instead of concave and that U_T and U_R will be positive. However, there is no obvious reason why a faculty member would necessarily want to be limited to positions on the technical performance constraint in the case where teaching and research yield positive utility, and a different constraint would seem to be more appropriate.

McKenzie himself notes the possible unrealism of his results implied by the assumption that T and R are not related to income. The consequences of introducing income into the analysis of faculty behavior can be demonstrated by maximizing a utility function $U(L, R, T, Y)$ subject to a constraint $Y = Y(T, R)$ linking income to time spent on research and teaching, and a time constraint $(T + R + L = \text{constant})$. The first-order conditions for this problem are:

$$\begin{aligned} U_Y - \lambda 1 &= 0, \\ U_L - \lambda 2 &= 0, \\ U_T > \lambda 1\ Y_T - \lambda 2 &= 0, \\ U_R > \lambda 1\ Y_R - \lambda 2 &= 0, \end{aligned}$$

where $\lambda 1$ is the marginal utility of income and $\lambda 2$ is the marginal utility of leisure. These conditions apply whether teaching and research yield positive or negative utility to a particular individual. The third and fourth conditions indicate that the optimal combination of T and R must satisfy the following condition:

$$U_T - \frac{U_R Y_T}{Y_R} = \lambda 2\left(1 - \frac{Y_T}{Y_R}\right).$$

From this condition it is clear that only if $\lambda 2 = 0$ or $Y_T/Y_R = 1$ will the optimal marginal rate of substitution between T and R equal the ratio of the marginal effects of teaching and research time on income. The optimal relationship between the MRS and ratio of marginal income effects of T and R can be seen more clearly by noting that the third and fourth first-order conditions also imply that

$$U_T - U_R = \lambda 1(Y_R - Y_T).$$

When Y_R exceeds Y_T, U_T will exceed U_R, and vice versa. As noted below, this relationship is the opposite of that predicted by a model of faculty behavior which ignores the time constraint.

These conclusions are very similar to those reached in our earlier treatment of McKenzie's model. In fact, if the marginal effects of T and R on income are independent of the allocation of a faculty member's time between teaching and research, the income constraint is linear in $R - T$ space, and the diagrammatics of the model are identical to those outlined in our previous discussion of McKenzie's model. Apart from the shape of the indifference curves, the major difference is that the optimal income level, and hence the position of the income constraint in $R - T$ space, is itself endogenously determined by faculty productivity, preferences, and the monetary rewards to the output of teaching and research time.

Nonlinear income constraints can readily be analyzed diagrammatically and will change the optimal allocation of time but not the optimal conditions relating to allocations of time between teaching and research. It is instructive, however, to note the precise conditions under which the income constraint will be linear and to link the analysis with the allegation that an incentive exists for faculty to engage in research rather than teaching. The form of the income constraint depends on the productivity of a faculty member's teaching and research time and the monetary rewards to units of teaching and research output. The slope of the constraint indicating different combinations of T and R compatible with a particular income level will equal the ratio of the marginal productivity of the individual's time in teaching and research, multiplied by the ratio of the monetary rewards to a unit of teaching and research output.[1] Given the rewards to units of teaching and research output, the slope of the income constraint will be constant only if the ratio of the marginal productivity of faculty time in teaching and research is independent of the allocation of his time between these activities. If the slope is constant, and greater or less than unity, this fact implies that the marginal income return to T or R will exceed the marginal return to the other activity at all allocations of the individual's time. This in turn implies that the maximum income level compatible with a particular amount of time spent on teaching and research can be obtained by specializing completely in the activity with the highest marginal return. Even in these circumstances, however, complete specialization will not be optimal, due to the effect of convex preferences for T and R as already explained in connection with the discussion of McKenzie's model. The same discussion also indicates that the amount of time spent on the activity with the highest marginal income return will be greater than the level predicted by a model of faculty behavior which ignores the time constraint.

---

[1] This is quite compatible with a situation in which a unit of time spent on research or teaching activities produces both research- and teaching-oriented outputs. This merely means that the ratio of the marginal physical productivity of a faculty member's time in teaching and research activities equals the ratio of the sum of the teaching and research outputs resulting from a unit of T and R, respectively. The inclusion of joint outputs from the use of faculty time does not imply that T and R are perfect substitutes in producing teaching and research outputs, however.

In the light of the preceding comments, the question arises as to why the marginal income return to research should be higher than the return to teaching for faculty in general, in view of the existence of wide differences in the relative ability of individual faculty in teaching and research. Given the rewards to units of teaching and research output, a sufficiently high ratio between the marginal productivity of faculty time in teaching and research can result in a situation in which *T* yields higher income than *R*. One possible explanation why this does not occur in general is failure to monitor the outputs of faculty time spent on teaching-related activities; output cannot be rewarded unless monitored, and it is the ratio of monitored outputs of *T* and *R* which determine the relative income returns to *T* and *R*.

Our analysis yields testable implications which indicate whether an income incentive to engage in teaching or research exists in practice for any particular faculty member. Since the marginal effects of *T* and *R* on income depend also on the marginal productivity of faculty time, it is impossible to infer their relative magnitude simply on the basis of the monetary rewards to teaching and research output.[2] However, our previous analysis demonstrates that if the marginal effect of *R* on income exceeds the effect of *T*, the optimal MRS between teaching and research will exceed unity; conversely, when the marginal income return to *T* exceeds the return to *R*, the optimal MRS will be less than unity. These implications are the opposite of those of a model which ignores the time constraint confronting faculty; such a model predicts that if the optimal MRS exceeds unity, the marginal income return to *T* is higher than the return to *R*.

DOUGLAS NEEDHAM

State University of New York, Brockport

References

McKenzie, Richard B. "The Economics of Reducing Faculty Teaching Loads." *J.P.E.* 80, no. 3 (May/June 1972): 617–19.

Siegfried, John J., and White, Kenneth J. "Financial Rewards to Research and Teaching: A Case Study of Academic Economists." *A.E.R.* 63, no. 2 (May 1973): 309–15. (*a*)

———. "Teaching and Publishing as Determinants of Academic Salaries." *J. Econ. Educ.* 4 (Spring 1973): 90–99. (*b*)

[2] Some empirical evidence on the magnitude of these rewards is contained in Siegfried and White (1973*a*, 1973*b*).

Assistant Professors Should Be Discriminated Against, or The Less Productive I Am, the More I Should Be Paid

In a recent note appearing in this *Journal* Tullock (1973) states that assistant professors should be discriminated against since their internal value (utility) of university output (published papers) is higher than that of senior staff members. That assistant professors have more incentive to publish than tenured faculty is probably an empirical fact. But to conclude that complementary inputs, such as secretaries, should be given to tenured rather than untenured faculty is to confuse marginal with total analysis.

If the university is in equilibrium, that is maximizing the number of published articles, then it may be quite true that to induce one more article from the faculty, more complementary inputs must be used with tenured, rather than with untenured, faculty. This argument implies that the university which wishes to maximize its output should devote it complementary resources to the younger faculty, given that the output is homogeneous.

All of this says nothing about the total amount of complementary factors that should be used with each type of professor in equilibrium. Depending on the production conditions (i.e., the marginal product schedule of the two types), either type of professor might receive more complementary inputs. According to Tullock's argument, the assistant professor has a high opportunity cost of not using his time for research, while the tenured faculty has a low opportunity cost. Hence, we may suppose that in any given time period, the assistant professor will devote more of that period to research than will the tenured member. If this is always so (e.g., the opportunity costs do not depend on the amount of time devoted to research), then more output can be obtained by freeing the assistant professor's time from mundane tasks such as typing than can be obtained from the tenured staff. In this case, all of the complementary inputs (secretaries, research assistants, etc.) should be given to the untenured staff, since by doing so the university will obtain more output than if it indulges any of the tenured faculty's demands for secretaries.

The world is not so simple as this. Published research is not homogeneous, and the opportunity cost of reasearch time probably does depend

---

The author is an assistant professor at Washington University and has benefited from discussions with Trout Rader, although Rader advised him, for productivity reasons, not to write this note.

on the amount of time spent on research. Hence, we find both types of professors supplied with some complementary inputs. Whether or not a university is maximizing its output with respect to the numbers of secretaries, etc., allotted to the different types of professors is an empirical question. But certainly there is no a priori argument that assistant professors should be discriminated against, if the university is interested in maximizing published research.

Tullock is correct, though, if the university is not interested in its total output, but in the total output per man. More specifically, Tullock is quite correct if the university's objective is to obtain at least one paper per given time period from each of its faculty. With this objective, his logic is infallible. Those with high opportunity costs of not devoting time to research will write papers anyway, while those with low opportunity costs of not doing research must be "bribed" to do it. But this means that the less productive a professor is, no matter what his rank, the more he should be paid (to be productive).

R. P. PARKS

Washington University

Reference

Tullock, G. "Universities Should Discriminate against Assistant Professors." *J.P.E.* 81, no. 5 (September/October 1973): 1256–57.

Puzzles and Problems

Inflation and the Terms of Trade

Professors Modigliani and Askari (question 730, *J.P.E.* 81 [January/February 1973]: 237–42) expressed surprise at their discovery of a very strong positive relationship between export prices and import prices of 19 countries over the period 1963–69. This relationship is summarized by a simple regression coefficient of .67, with the correlation coefficient being .57. Their surprise leads one to presume that their a priori reasoning led them to expect a lower regression coefficient, either zero, indicating no significant relationship between the prices, or less than zero, indicating a negative relationship.[1] I submit that their "puzzle" can be resolved by observing that there is reason to expect a positive relationship a priori and, further, by showing that when domestic prices are correctly taken into account, the data are consistent with such a priori reasoning. The crude hypothesis is that the terms of trade is a "real variable" which changes only slowly over time, whereas the price indices are dominated statistically by the inflation which occurred during the (relatively short) time period under consideration.

Consider the experiment put forward by Modigliani and Askari. In figure 1 herein, point *A* is the initial point (1963) for all 19 countries, 1963 being the year for which all price indices are set equal to 100. For any one country experiencing inflation, we would expect its observation in 1969 to lie northeast of point *A*, that is, prices of both imports and exports will have risen. If there are no changes in that country's terms of trade, we would observe the 1969 point for that country on the line 0*A* extended, the line of constant terms of trade with a slope of 1. If none of the countries has experienced any change in their terms of trade, all

---

I would like to thank Chin Lim for his assistance in preparing some of the calculations in this note.

[1] In a two-country, two-good world, one country's terms of trade are the inverse of the other's; hence, the observed relationship at any point in time is negative. In the multicountry, multigood case under consideration, trade is multilateral and the data are not disaggregated to identify bilateral terms of trade; there is no longer any case for expecting a negative relationship.

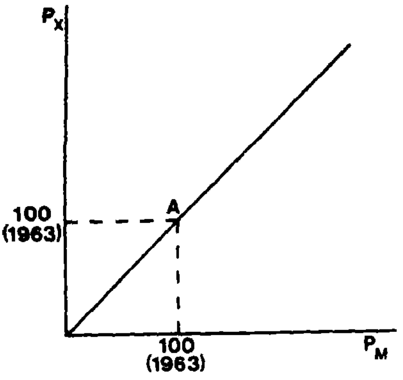

FIG. 1

19 observations in 1969 would lie on 0*A* extended, the points varying only by the differences in each country's rate of inflation. If this were the case, then the experiment conducted by Modigliani and Askari in order to estimate their equation (1) would give rise to a regression coefficient of 1. In fact, what happens is that most of the countries (all except for Great Britain) experience some change in their terms of trade, so that what we observe in 1969 is a scatter about the line 0*A* extended.[2] The computed regression coefficient is equal to .67.

If it is postulated that the terms of trade is a relative price which changes only slowly over time and that the 6-year period underlying the Modigliani-Askari experiment is indeed a short period for observing changes in such a variable, one would expect the scatter to be very close to the line 0*A* extended and would expect a computed coefficient of close to 1. This line of reasoning leads one to express surprise at the low value of the coefficient, not at the high value, which is what surprised them. (A further surprising implication of their coefficient of .67 is that it leads one to believe that countries which inflated faster—as measured by the distance of the 1969 observation from point *A*—tended to experience a decline in the relative prices of their exports. I examined this puzzling result directly by regressing the terms of trade on the inflation rate and by calculating a Spearman rank correlation coefficient. No significant relationship was found in either case.)

In order to better examine the hypothesis that inflation biased the regression results achieved by Modigliani and Askari, I conducted the following experiment. The data on export and import price indices were transformed by dividing each by a measure of overall change in the domestic price index. The only data apparently available across the entire set were on the consumer price index, so this was used for the deflation.[3] The variable denoting the price of exports relative to the overall consumer price index was then regressed on the variable denoting the price of imports relative to the overall consumer price index. These results are given in equation (1*):

---

[2] The diagram presented on p. 238 of their article contains some errors. Point 8 as it appears is wrong; it should lie due east of point 2. Point 9 was left off their diagram totally; it lies directly below point 14 but above the line of constant terms of trade.

$$\frac{PX_t}{P_{CPI_t}} = \underset{(.881)}{.091} + \underset{(7.19)}{.904}\,\frac{P_{Mt}}{P_{CPI_t}} \quad \begin{matrix} R^2 = .753 \\ \bar{R}^2 = .738 \end{matrix}\,. \qquad (1^*)$$

These results are rather startling. In terms of figure 1, what we are in principle doing is bringing the observations down to a scatter about point *A* by transforming each observation in their figure 1 along a ray toward the origin by the percentage increase in each country's consumer price index.[4] What we have computed is that the constant term is not significantly different from zero, and the regression coefficient on "real" import prices is not significantly different from 1. In addition, the explanatory power of the regression is increased immensely, with the new correlation coefficient being .86. Thus, to the extent that our transformed data truly reflect the changes in the "real" price of exports and imports, it is seen that inflation tends to generate an exaggerated variability in the terms of trade, and we are able to support the hypothesis that the terms of trade changes very slowly. Our evidence suggests that the "real" terms of trade for the 19 countries have not undergone any significant change during the 6-year period under consideration.

Parenthetically, I might report the results of the experiment with the transformed data when the observations for the countries which altered their exchange rates were deleted. These results are given in equation (2*):

$$\frac{Px_t}{P_{CPI_t}} = \underset{(.41)}{.06} + \underset{(5.11)}{.93}\,\frac{P_{Mt}}{P_{CPI_t}} \quad \bar{R}^2 = .68. \qquad (2^*)$$

These results are not significantly different from those reported in (1*), and hence we might offer this as evidence that exchange rate changes are not effective in altering the terms of trade.

DOUGLAS D. PURVIS

Queen's University, Kingston, Ontario

---

[3] These were the data used by Modigliani and Askari in their eq. (2). The transformation yields values in 1969 of the relative prices of exports and imports all substantially less than 1. This indicates that the prices of both exports and imports have been rising more slowly than the overall consumer price index, consistent with the stylized fact that the prices of services have been rising much more quickly than those of goods entering into international trade.

[4] In fact, the points would in general be southwest of point A (see n. 3 above).

Book Reviews

The World in Depression, 1929–1939. By CHARLES P. KINDLEBERGER. History of the World Economy in the Twentieth Century, vol. 4.
Berkeley and Los Angeles: University of California Press, 1973. Pp. 336. $10.00 (cloth); $3.45 (paper).

Charles Kindleberger's journalistic account of the world in depression after 1929 emphasizes the international scope of that disastrous decade. The structure is mostly chronological, each temporal chapter studded with brief vignettes devoted to developments in a particular time period in a number of industrial and developing countries. Since the basic material is familiar, a judgment on the book involves an assessment of the author's intellectual standards and of the validity of the interpretation he offers.

The author openly avows "intellectual idiosyncrasies" developed during his teaching career. These take the form of *obiter dicta* on substantive issues without supporting evidence and casual dismissal of opposing views. Illustrative examples will be cited in what follows. The author's interpretation is often inconsistent in detail and is loosely constructed in its broad outlines, combining selective strands in the conventional wisdom dating back to the 1930s on the significance of the depression. Alternative explanations are given short shrift, and hard questions are not dealt with. The book hardly qualifies as Everything You Wanted to Know about the Depression.

Kindleberger accepts that strand of the conventional wisdom which asserts that an unstable U.S. private sector must be stabilized by suitably offsetting fiscal policy of the government. His one novelty is to extend this assertion to an unstable world economic system "unless some country stabilized it, as Britain had done in the nineteenth century and up to 1913. In 1929, the British couldn't and the United States wouldn't" (p. 292).

Another strand in the conventional wisdom that Kindleberger accepts without offering a shred of evidence is that the depression decade was characterized by competitive devaluations and beggar-thy-neighbor policies. Had he examined the record, he could hardly have retained his belief in the validity of this proposition

The review is organized as follows: three sections discuss the origins, intensification, and resolution of the depression; a concluding section examines the implications that Kindleberger draws from the past for the current world economy.

Origins

Kindleberger includes as initial conditions for the depression the heritage of war, reparations, war debts, overvaluation of the pound and undervaluation of the franc, and the 1923 hyperinflation which led Germany to favor deflationary policies during the subsequent decade. He does not explain why or how this catalog fostered the depression. He attributes an important role to rising stocks and falling prices of world primary products after 1925 in propagating the depression, once started, although he later alludes to this factor as a "mild shock" to the system. According to Kindleberger, some countries did not adjust production of primary products to a small excess of supply over demand but instead stockpiled; other countries either followed suit or sold their primary products for what they would bring, sometimes depreciating their currencies to cushion the consequences at home but thereby forcing down prices abroad. (Kindleberger never explains how depreciation of the currency of a small country could force down prices abroad.) He defines this development as a structural deflation of the system.

Up to the middle of 1928, the deflation was papered over by "the abundance of credit," though apparently not a sufficient abundance to prevent the price declines. What changed the picture, according to Kindleberger, was the New York stock market boom after mid-1928 and the subsequent crash in October 1929. The boom deprived developing countries of capital, which was diverted to the stock market; the crash reduced exports of these countries as a result of a liquidity crisis in the United States, which was communicated to commodity markets and durable-goods industries. By the end of 1929, the feedback led to a sharp decline in U.S. exports. Kindleberger thus regards the crash as an important part of the deflationary mechanism rather than a superficial phenomenon.

"All this time, business was slipping" (p. 116), but Kindleberger apparently attaches no special importance to the upper turning points reached well before the New York stock market crash in the countries he refers to: Germany, the United States, Great Britain, Belgium, Canada, South Africa. In this context he treats the United States as just another country with a business decline. This is wholly inconsistent with his stress on U.S. abstention from the role of a leader as the explanation of the worldwide character of the depression. Surely a business contraction in the country assigned the role of a leader— and accounting for an estimated 40 percent of the world's output—is more critical to the world than contraction in other countries.

Kindleberger is not sympathetic to the view that the downturn in the United States was produced by overexpansion of investment as a result of a mechanical connection between multiplier and accelerator or of other more deep seated factors. The only real factor he gives any role is a decline in housing construction from 1925 to 1929. He dismisses out of hand a monetary explanation with one of his characteristic *obiter dicta* and without even a pretense of examining the evidence: "There would doubtless have been a recession or depression with perfect monetary policy, or a money supply which grew in the United States at *some optimal* pace" (p. 20).

I am led to infer that Kindleberger believes that the contraction in the United States was initiated by a decline in housing; that it started as a mild contraction and was transformed into a depression by the stock market crash. If this is correct, it implies that he regards the stock market, not real factors, as the source of the alleged instability of the private sector.

Perhaps because of his "prejudice against flexible rates" (p. 15), dating from his student days, Kindleberger offers no criticism of the operation of the gold standard. He refers to the requirements of symmetric behavior by the world economy: "Gold losses for one country would be deflationary, but gain for the recipient country would yield offsetting expansion" (p. 292). Yet, surprisingly, he does not discuss the failure of the United States to behave symmetrically. The Federal Reserve sterilized much of the gold inflow into the United States after 1923, preventing an expansionary effect on the stock of money and thence on prices, such as would have occurred under the prewar gold standard. Instead the system sought, and to a large measure achieved, stable economic growth with falling wholesale prices. This achievement was largely at the expense of economic stability in Great Britain and the peripheral countries tied to sterling. Granted that Britain returned to gold in 1925 at a parity that overvalued sterling, rising prices in the United States and the accompanying reduction in gold inflows would have significantly eased pressure on sterling. Similarly, France, after returning to the gold standard in 1928 at a parity that undervalued the franc, also proceeded to sterilize gold, further weakening Britain's position.

Monetary policy and developments readily account not only for the initial conditions Kindleberger cites, but also for the onset of the contraction in the United States and for the severity of the first 14 months of the contraction. The onset of the contraction in the United States (in August 1929, not in June, as Kindleberger states) reflected a monetary policy in 1928 that was too restrictive to permit stable economic growth, yet not restrictive enough to curb the stock market boom. The stock of money did not grow at all during the 16 months preceding the peak—a longer period of no growth than preceded any prior or subsequent business expansion.

Kindleberger has misread or not read the discussion in *A Monetary History of the United States, 1867–1960* (Friedman and Schwartz 1963) of the first 14 months of the contraction in the United States, which describes the decline of 2.6 percent of the money stock from August 1929 to October 1930 as "a larger decline than during the whole of all but four preceding reference cycle contractions . . . and all the exceptions are contractions that were exceptionally severe by other indications as well" (pp. 307–8). "Even if the contraction had come to an end in late 1930 or 1931, as it might have done in the absence of the monetary collapse that was to ensue, it would have ranked as one of the more severe contractions on record" (p. 306). On Kindleberger's misreading, "Friedman and Schwartz state that the stock of money started to decline in October 1930, after having gone sideways. This is difficult to accept on their figures" (p. 137, n. 19).

He then asserts: "There is no evident mechanism by which firms and financial intermediaries exchanging government bonds for money could have stopped it [the decline in commodity prices]. . . . The inefficacy of monetary policy at this stage—which is not to say that it was not badly handled later—is illustrated by the fact that yields on bonds which had been moving in concert went separate ways. That on government and corporate Aaa bonds declined; Baa bonds issued by firms of less certain financial standing went from under 6 to over 9 percent. It was not sufficient to make money abundant and cheap; one also had to improve credit-worthiness by improving the outlook" (pp. 137–38).

These sentences are a compendium of items in the conventional wisdom for which there is no empirical evidence. (1) One such item is that the effect of a given change in the quantity of money depends in important measure on how the change is produced; specifically, that monetization of government bonds has a different effect on prices than commercial loans by banks to private borrowers. The three studies of this issue that can be cited (Auerbach 1969; Bordo 1972; Cagan 1972) conclude that the first-round effect of the source of money creation is far outweighed by subsequent rounds of portfolio adjustments. (2) A second item in the conventional wisdom is that the spread of interest rates among high-grade and low-grade assets proves the inefficacy of monetary policy. It does no such thing. It simply reflects a shift in demand by banks and others from more to less risky investments, as a result of the Federal Reserve's massive withdrawal of credit from the banking system during the period from August 1929 to October 1930. (3) A third item is that credit was abundant and cheap during the initial phase of the depression. It was not abundant, as the unprecedented size of the decline in the money stock attests. And Kindleberger himself acknowledges at another point that credit was not cheap (p. 144). (4) A final point is that credit worthiness is a variable that can be manipulated independently from the "quantity" of credit and its price—a proposition which in Kindleberger's presentation becomes simply a restatement of his belief in the impotence of monetary policy.

Intensification

Kindleberger traces the intensification of the depression to repercussions of the initial contraction on the economies of the countries producing primary products, not to developments in the United States. He emphasizes the halt in international lending to peripheral countries—he is unsure whether as a result of a decline in demand or a decline in supply because of dubious prospects for debt service in a world of falling prices. Given their limited gold and foreign exchange reserves, he believes that they were forced to sell their primary products for whatever they would bring. Tariff increases and quota restrictions exacerbated the problems of the primary producing countries and reduced world trade. Kindleberger does not note that countries on fixed exchange rates applied import controls, not countries that devalued. On his reading, falling security and commodity prices, for which he assigns no special responsibility to the United States, made banking systems everywhere vulnerable and led to the financial crisis of 1931.

Kindleberger mentions that, in May 1931, "The European press kept insisting that United States policies sucked up gold and undermined European financial systems" (p. 152), but only to note that President Hoover did not agree. He does not himself point out that the European press was right and Hoover was wrong. From 1929 to 1931, the United States had an inflow of gold of over half a billion dollars. Far from expanding the U.S. money stock in response to the gold inflow, the Federal Reserve engineered a decline, thereby preventing any domestic price and income adjustment and imposing the whole burden of adjustment on countries losing gold.

The effects of U.S. monetary policy after 1929 first became severe in those countries that had returned to gold with the smallest gold reserves and with financial structures weakened seriously by World War I—Austria, Germany, and Hungary. The Federal Reserve participated in the extension of modest short-term credits to the central banks of these countries but, given the basic pressure on them of deflation in the United States and the fixed exchange rate link which bound them to gold, such assistance was at best a temporary palliative. Frantic efforts to save sterling, which next came under pressure, were likewise futile. Kindleberger is silent on this source of the spreading financial crisis in Europe in 1931.

Britain did not undertake monetary expansion until February 1932 and so got no immediate relief from deflation by abandoning the gold standard. Deflation continued even longer in the gold-standard countries, including France, Germany, and the United States. Germany followed a deflationary domestic policy partly in the belief that it would bring the end of reparations by demonstrating the impossibility of payment, partly in response to the hyperinflation syndrome. For the United States, Kindleberger accepts the (discredited) Federal Reserve view that by the end of 1931 a shortage of free gold precluded open market purchases, but sees no contradiction when he reports that congressional pressure finally induced the Open Market Committee to expand the modest purchase program it voted after the passage of the Glass-Steagall Act. (He confuses the vote taken on adoption of the modest program with the vote on the expansion of the program.)

Kindleberger argues that the large-scale open market purchases between

March and July 1932 did little to relieve the squeeze on the economy outside financial markets because country banks still had sizable net borrowed reserves by the end of 1932, even though New York and Chicago banks had excess reserves and reserve city banks were virtually out of debt. He fails to note that country banks reduced their debt by one-half from the start to the end of the program and would doubtless have reduced it still further if the program had been continued throughout the balance of the year. Kindleberger's negative attitude toward monetary policy leads him into outright error when he reports that commodity prices continued to decline after March 1932, when the open market purchases were initiated, even though his own table 11 shows that 12 out of 14 commodity prices rose from March to September 1932. He contradicts himself at a later point when he notes that the depression appeared to be bottoming out in the spring of 1932, but is agnostic about "whether one ascribes this to the $1,000 million of open-market operations, to the end of appreciation [of the dollar], or to the multiplier-accelerator turning point which comes when gross investment hits the bottom" (p. 194).

Resolution

Kindleberger invokes a real factor to account for the end of the contraction in the United States: when gross investment was close to zero at the end of 1932, the system was ready for an upturn. The upturn, however, was aborted: "It was first necessary [why?] to have an apocalyptic climax in bank failures" (p. 194). Presumably, thereafter expansion ensued, but the reader is differently informed at a later point when Kindleberger refers to "the accentuation of the depression which came with the transfer of the presidency from Hoover to Roosevelt (occurring after the money supply had been greatly enlarged); and the still more significant . . . transfer of leadership in the world economy from Whitehall to the White House" (p. 300). This statement is mystifying. The revival of business from the cyclical trough in March 1933 was erratic and uneven, but the relapse in the second half of 1933 cannot accurately be described as an accentuation of the depression. Industrial production and personal income remained at higher levels than their trough values throughout the remainder of the year, as Kindleberger himself suggests at another point (p. 233). Further, the money stock was not greatly enlarged. On the contrary, both M_1 and M_2 were essentially unchanged from their trough values in April 1933 until the end of the year.

Kindleberger argues that two means were available to achieve world recovery: (1) simultaneous programs of government spending in all countries; (2) simultaneous devaluations to create gold profits which would be available for spending. Conspicuous by its absence is any discussion of monetary expansion and the liberation of monetary policy made possible by floating exchange rates in the 1930s, a regime he deplores. Kindleberger mentions the deflation that persisted in the gold-bloc countries after recovery was underway in the countries that had devalued but does not allude to the monetary implications of the contrast. Yet, in discussing the reason industrial countries were relatively immune to the spread of the U.S. 1937–38 contraction, Kindleberger cites their easy monetary policy and substantial spending for public works (p. 278). Why was monetary policy important in this contraction and not in the preceding one? Did floating exchange rates play no part in the experience?

Unlike 1929–33, the 1937–38 contraction in Kindleberger's view originated in the United States. What was distinctive about 1937–38 that was not true of 1929–33? The reader is not informed. Three factors are usually cited as contributing to the contraction: a reversal of unsustainable inventory accumulation; tighter money as a consequence of increases in reserve requirements; and a sharp change in fiscal policy. Kindleberger emphasizes only the first. He regards recovery as a result of the final acceptance of the Keynesian doctrine of spending for stability. He explains the duration of the depression in the United States by the resistance for 7 years of this doctrine by both Hoover and Roosevelt. For the rest of the world he has a different explanation. According to Kindleberger, the world economy was unstable for so long because there was no single leader to stabilize it, the British being impotent and the United States shirking the responsibility. Three tasks a leader would have assumed were left undone: (1) maintenance of a relatively open market for "distress goods," that is, imports; (2) provision of countercyclical long-term lending; (3) discounting in crises.

There is a good deal of ambiguity in this conclusion. How does a leader operate—through governmental or private institutions? Is it market forces that impel a leader to act as Kindleberger expects it, or is it a moral force?

As for the functions he assigns the leader, Kindleberger couples the first one on his list—to keep the import market open during business depression—with a denunciation of competitive depreciation (p. 294). As already noted, tariff walls were erected in a vain effort to preserve existing exchange parities, but Kindleberger does not learn the lesson the experience teaches: floating exchange rates in the depression decade would have expanded world trade and employment, whereas futile attempts to preserve overvalued currency rates reduced world trade and employment. As for the alleged exchange war in the 1930s, there is no evidence that it occurred. External circumstances forced devaluations, not the deliberate undercutting of a competitor's exchange rate in order to increase a trade surplus. The devaluation forced on Britain in 1931 was not followed by greater devaluations among its trading partners (Rolfe and Burtle, 1973). On the contrary, the gold-bloc countries, of which France was the leader, did not devalue at all in response. Japan's devaluation exceeded Britain's but was dictated by domestic expansionary policy rather than by competitive trade reasons. When France finally devalued in 1935, other countries in the gold bloc did not devalue to a greater extent, nor did the sterling-area countries respond by a further devaluation. If one country alone devalues under the force of external pressure, even if it is followed by the devaluation of closely tied currencies, what has happened is needed adjustment—not competitive devaluation. Finally, if peripheral countries devalue, the impact on world trade is minimal.

With respect to the second function Kindleberger assigns the leader—countercyclical long-term lending abroad—his description of pre–World War I experience is simplistic, and its applicability to 1929–39 is suspect. As is well known, inverse movements of British home and overseas investment characterized only longer, not shorter cyclical swings (there were 32 positive, nine inverse annual movements in the two series, 1871–1913). But, even granting Kindleberger's point for the longer run, is it true that inverse movements of the series made an essential contribution to world stability? What in fact happened were bursts of development in particular areas. Even when net foreign investment was low, large-scale geographical shifts in investment occurred, benefiting some areas at the expense of others. For particular areas, since the rate of lending was uneven, development was marked by crises. Whether or not this pattern should be designated as world stabilizing, it is clear that the flow of capital before World War I was almost entirely a response to market influences, and the international demand for capital was directed largely to railway construction. What relevance does this experience have for 1929–39? Kindleberger makes much of the fact that for the interwar period the longer-run relationship between net capital formation in the United States as estimated in the national accounts and long-term capital transactions in the balance of payments was positive and not inverse (though annual movements were about equally divided between positive and inverse movements). However, had the United States not been plunged into depression, and had both domestic capital formation and capital export remained high as in the 1920s, why should such a positive pattern be regarded as harmful to the world?

The third function of a leader, according to Kindleberger, is discounting in crises. He does not recognize that crises are created by misalignment of exchange parities, and he does not perceive that discounting is a palliative that ultimately proves ineffective. Floating exchange rates obviate the need for this function.

Insofar as there was a world cycle in the 1930s, my conclusion is that its source was in the United States, not because the United States was not behaving according to Kindleberger's norms, but because the size of the U.S. economy and the violence of its fluctuations dominated the scene for countries that were tied to it by fixed exchange rates.

Relevance to the 1970s.

Under this rubric, Kindleberger argues that U.S. economic leadership is beginning to slip. He foresees three possible stable and three unstable outcomes. The stable outcomes are (*a*) a revival of U.S. leadership after exchange controls and protectionism have been reversed; (*b*) the replacement of U.S. by European leadership; (*c*) the creation of international leadership through a world central bank, a world capital market, and an effective GATT. The unstable outcomes are (*a*) U.S. and European competitive efforts at leadership; the United States and Europe each (*b*) either unable or unwilling to lead; or (*c*) stalemated by the other's right to veto programs of stability.

The gift of prophecy is denied to most of us. The problems of the 1970s are rather different now than they were thought to be by Kindleberger only a year ago. Inflation is now the world's chief economic problem. If the floating exchange rate system is retained, no leader is needed to curb inflation. Each country can do so on its own. Other problems, such as how to deal with the cartel determining the world supply and price of petroleum, may require cooperative action by the consuming nations, but it is hard to link the solution to the lessons of the depression decade.

Kindleberger's' account of the past is unfortunately no more satisfactory than his forecast for the future.

ANNA J. SCHWARTZ

National Bureau of Economic Research

References

Auerbach, Robert. "The Income Effects of the Government Deficit." Ph.D. dissertation, Univ. Chicago, 1969.

Bordo, Michael D. "The Effects of the Sources of Change in the Money Supply on the Level of Economic Activity: An Historical Essay." Ph.D. dissertation, Univ. Chicago, 1972.

Cagan, Phillip. *The Channels of Monetary Effects on Interest Rates.* New York: Nat. Bur. Econ. Res., 1972.

Friedman, Milton, and Schwartz, Anna J. *A Monetary History of the United States, 1867-1960.* Princeton, N.J.: Princeton Univ. Press (for Nat. Bur. Econ. Res.), 1963.

Rolfe, Sidney E., and Burtle, James. *The Great Wheel: The World Monetary System, A Reinterpretation.* New York: Quadrangle, 1973.

Inflation, Unemployment, and Poverty. By JOHN L. PALMER.
Lexington, Mass.: Lexington Books, 1973. Pp. xiii+170. $12.50.

This volume reports the results obtained by the author in two rather distinct areas of economics, namely the income distribution effects of inflation and a reduced-form view of the Phillips Curve. Both of these areas relate somewhat to the traditional inflation literature; therefore, the author saw fit to combine the two studies into one book. Palmer would have been much better off, however, producing two articles of significant worth, rather than one book, as this arrangement would have allowed the rather substantive propositions of the author much broader exposure than will be possible from this volume.

The first of these articles could have been the treatment of the impact of post-World War II inflation on the position of the poor. The empirical evidence presented does not support the traditional line of the profession that inflation hurts the poor and aged. Rather, by using the *Survey of Consumer Expenditures, 1960-61* and *Survey of Financial Characteristics of Consumers, 1961,* Palmer presents an impressive array of data to counter our classroom folly. The author uses a three-stage analysis here, examining the relative price movements for different income groups, the relative income changes over the period, and the change in net worth. By using the expenditure survey, a disaggregated annual price index is presented for various income and demographic groups, that is, poor, near poor, aged poor, urban white poor, urban nonwhite poor, rural poor, and—finally—the wealthy. The conclusion reached here is stated quite succinctly by the author: "It seems to us that little, if any, significance can be attached to the differences. . . . The strongest statement that could be justified is . . . that cost-of-living effects of inflation in the 1960s have been less adverse for the poor than for the non-poor" (p. 65). Using the *Financial Characteristics Survey,* he then proceeds to the impact of inflation upon income and wealth. On income he presents strong evidence to suggest that the fixed-nominal-income characterizations of the poor are dubious. Here two often-considered sources of income are examined —return on fixed assets and fixed-income instruments. Considering the first of these, rent is found to be uniformly distributed, or sufficiently so to negate any real distribution effects across income levels. Further, less than 10 percent of the aged poor receive income from pensions, annuities, and other sources that are fixed in nominal terms. On the other hand, the real value of various forms of public assistance, a very important income source for the aged poor, actually grew over the period, at times at a greater rate than per capita disposable income. For the nonaged poor, earned income—their major source of income—benefits from increases in economic activity, even if such expansion was at the expense of inflation. This further suggests beneficial aspects of inflation (relatively) on the poor. Finally, if we turn to wealth balances, it is demonstrated, under rather arbitrary assumptions that appear to be quite conservative, "that taken as a group, the poor, and particularly the aged poor, do not suffer substantial losses in their real wealth due to inflation" (p. 84). There are two causes of concern, however, in connection with this result. First, the data on net worth are too crude and heterogeneous in classification. Much of the insight into the distributional question is lost by such broad grouping, as little attention can be given to implied balance-sheet constraints, and resultant benefits to sectors issuing these liabilities. The second point of concern is the assumption that this composite internal debt is then assumed to have the same rate of return, that is, zero. The author attempts to remedy this by supporting his results with more precise estimates published elsewhere that essentially verify the conclusions offered here. Yet even this does not eliminate the skepticism that develops over the robustness of the results presented. In all, however, this part of the study is worthy of note, and the subject area is worthy of further research.

The second of the proposed articles could have been the author's treatment of inflation and unemployment, and the so-called trade-off. The author's position here is that the "traditional Phillips Curve analysis" inaccurately portrays causality between the rate of change of wages and unemployment in its usual empirical formulation. Palmer prefers to consider wage and price change a function (in a causal sense) of their expected values and unemployment a function (in a causal sense, again) of price and wage change and their expected values. The point made by the author is that prices, wages, and unemployment are endogenous to the system, and the early empirical analysis inappropriately specified unemployment as an independent variable. Using the equilibrium condition that expected rates equal actual rates, and a partial adjustment mechanism using both a geometrically decaying lag formulation and the Feldstein-Eckstein expectations series, he conducts empirical estimation on quarterly data, fourth quarter of 1959 to fourth quarter of 1969. The success of the estimation on price change using expected values obtained from lagged actuals is not surprising, or, in my opinion terribly informative. The results on unemployment, while relatively successful, are uniformly less accurate, in an R^2 sense, and exhibit low Durbin-Watson statistics. Next, the author conducts a two-stage, least-squares estimation procedure on expected price change and on expected wage change using the monetary base and full-employment surplus as determinants of these expectations. Here the results are substantially the same as the ordinary least-squares estimates. There is much to quarrel with in this analysis, however, on both theoretical and ideological grounds. The basic point of this section of the study is the need for a general-equilibrium approach to wage and price change and unemployment, that is, a call for reduced-form equations determining these variables. This, I think, is to be lauded, although it appears that we are beating a straw man, at least in terms of the present literature. But the author then refuses to use a complete reduced form, eliminating the price-expectations equations, and to argue that real and monetary variables determine prices, wages, and unemployment. Instead, at a rather arbitrary point, it is assumed that monetary and fiscal policy, as a set, affect only price expectations, and price expectations affect actual price and employment movements. This point appears no less arbitrary than the partial analysis of earlier Phillips Curve work. Further, the policy variables themselves are not exogenous but are determined by the real variables in the system. This fact leads the reader to doubt the appropriateness of the specification—along with the author, who interjects: "We are far from convinced that this assumption is appropriate" (p. 103).

However, let us return to the volume as it stands. The study itself contains three relatively distinct parts. The first four chapters contain a rather straight forward summary of the inflation literature. The next five contain the impact of inflation on the poor as discussed above, and chapter 10 contains the Phillips Curve empirical work. Those interested in the individual sections of the work would do well to refer to it. The whole, however, is less than the potential sum of its parts.

ANTHONY M. SANTOMERO

University of Pennsylvania

THE WORD'S OUT—

ECONOMICS: Principles, Problems, and Policies, Sixth Edition

Campbell R. McConnell, University of Nebraska

January, 1975 960 pages (tent.) $12.50 (tent.)

Building on the time-proven benefits of its previous editions, the new Sixth has been expanded to include new and additional material such as: an entire chapter on the Phillips Curve; a new chapter on the pros and cons of radical economic theory; "Last Word" essays following each chapter which have been drawn from varied sources; and new and important material on the macroeconomics of international trade.

Available with:

Instructor's Manual. Test File.

Overhead Transparency Set, 1975, $210.00 (tent.). **Transparency Masters,** 1975, $15.00 (tent.). **ECONOMIC ISSUES,** A Book of Readings, Fifth Edition. *Edited by* Campbell R. McConnell, University of Nebraska, 1975, 480 pages (tent.), $6.50 (tent.) soft. **STUDY GUIDE,** Sixth Edition, Robert C. Bingham, Kent State University, 1975, 368 pages (tent.), $4.50 (tent.). **ECONOMIC CONCEPTS:** A Programmed Approach, Fourth Edition, Robert C. Bingham, Kent State University, 1975, 400 pages (tent.), $5.50 (tent.). **ECONOMICS, MATHEMATICALLY SPEAKING,** Robert C. Bingham, Kent State University, 1972, 424 pages, $5.95. **PERSONALIZED SYSTEM OF INSTRUCTION** for use with **ECONOMICS,** Sixth Edition, Jerry Petr, University of Nebraska, 1975, 256 pages, 128 test pages (tent.), free to adopters.

URBAN ECONOMICS AND POLICY ANALYSIS

Robert L. Bish, University of Southern California and **Hugh O. Nourse,** University of Missouri at St. Louis

January, 1975 384 pages (tent.) $10.95 (tent.)

The most distinctive feature of this new text is its integration of location theory and the theory of externalities, public goods, and collective action, to better analyze urban problems and policy alternatives.

A HISTORY OF ECONOMIC THEORY AND METHOD

Robert B. Ekelund, Jr., Texas A&M University and **Robert F. Hebert,** Auburn University

January, 1975 475 pages $11.95 (tent.)

This systematic and detailed treatment of the history of economic thought emphasizes how the theory and methods of present and past economists combined to form contemporary economics.

"Rational" Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule

---

Thomas J. Sargent and Neil Wallace

University of Minnesota

Alternative monetary policies are analyzed in an ad hoc macroeconomic model in which the public's expectations about prices are rational. The ad hoc model is one in which there is long-run neutrality, since it incorporates the aggregate supply schedule proposed by Lucas. Following Poole, the paper studies whether pegging the interest rate or pegging the money supply period by period minimizes an ad hoc quadratic loss function. It turns out that the probability distribution of output—dispersion as well as mean—is independent of the particular deterministic money supply rule in effect, and that under an interest rate rule the price level is indeterminate.

This paper analyzes the effects of alternative ways of conducting monetary policy within the confines of an ad hoc macroeconomic model. By ad hoc we mean that the model is not derived from a consistent set of assumptions about individuals' and firms' objective functions and the information available to them. Despite this deplorable feature of the model, it closely resembles the macroeconomic models currently in use, which is our excuse for studying it. Following Poole (1970), we compare two alternative strategies available to the monetary authority. One is to peg the interest rate period by period, letting the supply of money be whatever it must be to satisfy the demand for it. The other is to set the money supply period by period, accepting whatever interest rate equilibrates the system. We study the effects of such policies for two versions of the model: an autoregressive version in which the public's expectations are assumed formed via fixed autoregressive schemes on the variables being forecast, and a rational-expectations version in which the public's expectations are

Work on this paper was supported by the Federal Reserve Bank of Minneapolis, which does not necessarily endorse the conclusions. Robert Barro, Milton Friedman, John Kareken, and Robert E. Lucas made useful comments on an earlier version of the paper.

[*Journal of Political Economy*, 1975, vol. 83, no. 2]

© 1975 by The University of Chicago. All rights reserved.

assumed equal to objective (mathematical) expectations that depend on, among other things, what is known about the rules governing monetary and fiscal policy (see Muth 1961).

The two versions have radically different policy implications. In the rational-expectations version, (*a*) the probability distribution of output is independent of the deterministic money supply rule in effect, (*b*) if the loss function includes quadratic terms in the price level, then the optimal deterministic money supply rule is that which equates the expected value of next period's price level to the target value, and (*c*) a unique equilibrium price level does not exist if the monetary authority pegs the interest rate period by period, regardless of how its value varies from period to period. None of these results emerges from the autoregressive version. It, instead, exhibits all the usual exploitable tradeoffs between output and inflation, implies that minimization of the above loss function is a well-defined nontrivial dynamic problem giving rise to a unique optimal deterministic feedback rule either for the money supply or for the interest rate, and has a unique period-by-period equilibrium if the interest rate is pegged. Thus, in the autoregressive version of the model, which in principle is merely a variant of Poole's model, whether an interest rate feedback rule or a money supply feedback rule is superior depends, just as Poole asserted, on most of the parameters of the model including the covariance matrix of the disturbances.

In the rational-expectations version of the model, one deterministic money supply rule is as good as any other, insofar as concerns the probability distribution of real output. In this weak sense, an *X* percent growth rule for the money supply is optimal in this model, from the point of view of minimizing the variance of real output. Thus, switching from the assumption of autoregressive expectations to that of rational expectations has drastic policy implications. In particular, making that change transforms the model in which following Friedman's *X* percent growth rule would in general be foolish into one in which such a rule can be defended as being the best the authority can do.

1. The Ad Hoc Model

We assume a structure described by the following equations:[1]

aggregate supply schedule:

$$y_t = a_1 k_{t-1} + a_2(p_t - {}_t p^*_{t-1}) + u_{1t}, \qquad a_i > 0, i = 1, 2; \tag{1}$$

aggregate demand schedule or "*IS*" curve:

$$\begin{aligned} y_t = {} & b_1 k_{t-1} + b_2[r_t - ({}_{t+1}p^*_{t-1} - {}_t p^*_{t-1})] \\ & + b_3 Z_t + u_{2t}, \qquad b_1 > 0, b_2 < 0; \end{aligned} \tag{2}$$

---

[1] The structure closely resembles the model used by Sargent (1973).

portfolio balance or "*LM*" schedule:

$$m_t = p_t + c_1 y_t + c_2 r_t + u_{3t}, \qquad c_1 > 0, c_2 < 0; \tag{3}$$

determination of productive capacity:

$$k_t = d_1 k_{t-1} + d_2[r_t - ({}_{t+1}p^*_{t-1} - {}_tp^*_{t-1})] + d_3 Z_t + u_{4t}, \quad d_2 < 0; \tag{4}$$

evolution of the exogenous variables:

$$\begin{aligned} Z_t &= \sum_{j=1}^{q} \rho_j Z_{t-j} + \xi_t, \\ u_{it} &= \sum_{j=1}^{q} \rho_{ij} u_{i,t-j} + \xi_{i,t}. \end{aligned} \tag{5}$$

Here y_t, p_t, and m_t are the natural logarithms of output, the price level, and the money supply, respectively; r_t is the nominal rate of interest itself (not its logarithm); while Z_t is the vector of exogenous variables. The variable ${}_{t+i}p^*_{t-j}$ is the public's psychological expectation of the log of the price level to prevail at $t + i$, the expectation being held as of the end of period $t - j$. The variable k_{t-1} is a measure of productive capacity, such as the logarithm of the stock of capital or labor or some linear combination of the logarithms of those stocks at the end of period $t - 1$.

Equation (1) is an aggregate supply schedule relating output directly to productive capacity and the gap between the current price level and the public's prior expectation of the current price level. Unexpected increases in the price level thus boost aggregate supply, the reason being that suppliers of labor and goods mistakenly interpret surprise increases in the aggregate price level as increases in the relative prices of the labor and goods they are supplying. This happens because suppliers receive information about the prices of their own goods faster than they receive information about the aggregate price level. This is the kind of aggregate supply schedule that Lucas (1973) has used to explain the inverse correlation between observed inflation and unemployment depicted by the Phillips curve.

Equation (2) is an aggregate demand or "*IS*" schedule showing the dependence of aggregate demand on the real rate of interest and capacity, a measure of wealth. The real rate of interest equals the nominal rate r_t minus the rate of inflation between t and $t + 1$ expected by the public as of the end of period $t - 1$, namely, ${}_{t+1}p^*_{t-1} - {}_tp^*_{t-1}$. The rate r_t is assumed to be the yield to maturity on a one-period bond. Aggregate demand also depends on a vector of exogenous variables Z_t which includes government expenditures and tax rates.

Equation (3) summarizes the condition for portfolio balance. Owners of bonds and equities (assumed to be viewed as perfect substitutes for one another) are satisfied with the division of their portfolios between money, on the one hand, and bonds and equities, on the other hand, when equation (3) is satisfied. Equation (3) posits that the demand for real balances depends directly on real income and inversely on the nominal rate of interest.

Equation (4) determines productive capacity for the next period, while equation (5) describes autoregressive processes for the exogenous variables. The ζ's, which are sometimes called the "innovations" in the Z and u processes, are serially uncorrelated random variables.

To complete the model, we need equations describing ${}_{t+1}p^*_{t-1}$ and ${}_tp^*_{t-1}$. Adding those equations to (1)–(5) then results in a system capable of determining the evolution over time of y_t, p_t, r_t, ${}_{t+1}p^*_{t-1}$, and ${}_tp^*_{t-1}$ and k_t.

2. The Stabilization Problem

In order to discuss policy within the context of an ad hoc model, we must adopt an ad hoc loss function. The most familiar such function is the quadratic loss function

$$L = E_0 \sum_{t=1}^{\infty} \delta^{t-1}\left[(y_t, p_t)K(y_t, p_t)' + (y_t, p_y)(K_1, K_2)' + \frac{K_1^2}{4} + \frac{K_2^2}{4}\right],$$

where K is diagonal with elements $K_{ii} > 0$, $i = 1, 2$; K_1 and K_2 are parameters; and $0 < \delta < 1$. This function is separately quadratic in y and p and implies that $L = 0$, its lower bound, at particular constant values of y and p, the target values $-K_1/2K_{11}$ for y and $-K_2/2K_{22}$ for p. This function is easy to work with because it is quadratic, additive over time, and stationary in that the function of y and p whose expectation is to be minimized does not depend on t.

To minimize L, the monetary authority compares two strategies. The first is to peg r_t via a deterministic linear feedback rule

$$r_t = G\theta^*_{t-1}, \tag{6}$$

where θ^*_t represents the set of current and past values of all of the endogenous and exogenous variables in the system as of the end of period t, and G is a vector of parameters conformable to θ^*_{t-1}. The monetary authority chooses the parameters in G to minimize L. It must then compare the minimum loss associated with an interest rate rule having those G's with the loss associated with the best money supply feedback rule of the form

$$m_t = H\theta^*_{t-1}. \tag{7}$$

Whichever rule delivers the lower loss is the one that should be used.

3. The Autoregressive Expectations Version

Here we assume that the psychological expectations ${}_tp^*_{t-1}$ and ${}_{t+1}p^*_{t-1}$ are governed by the distributed-lag or "adaptive" schemes

$$ {}_{t+1}p_t^* = \sum_{i=0}^{q} v_{1i} p_{t-i}, \tag{8} $$

$$ {}_{t+2}p_t^* = \sum_{i=0}^{q} v_{2i} p_{t-i}, \tag{9} $$

where the v_{1i}'s and v_{2i}'s are fixed numbers. Given that the money supply is used as the monetary instrument, the system formed by equations (1)–(5), (8), and (9) can be reduced to a difference equation of the form

$$ Y_{1t} = \sum_{i=1}^{q'} A_i Y_{1t-i} + \sum_{i=0}^{q'} B_i m_{t-i} + \phi_{1t}, \tag{10} $$

where $Y'_{1t} = (y_t, p_t, r_t, k_{t-1}, Z_t)$ and ϕ_{1t} is a vector of serially uncorrelated random variables, the components of which are functions of the ξ_t's in equations (5). The A_i's are vectors conformable with Y_{1t} and the B_i's are scalars; both the A_i's and B_i's depend on the parameters of equations (1)–(5), (8), and (9). To find the best money-supply feedback rule, the monetary authority chooses the parameters H of the rule (7) to minimize the loss L subject to (10). Where loss is quadratic and the model is linear with known coefficients, rules of the linear form of (7) are known to be optimal.[2]

To find the optimal interest rate rule, the system formed by equations (1)–(5), (8), and (9) is written as

$$ Y_{2t} = \sum_{i=1}^{q'} C_i Y_{2t-i} + \sum_{i=0}^{q'} D_i r_{t-i} + \phi_{2t}, \tag{11} $$

where $Y'_{2t} = (y_t, p_t, m_t, k_{t-1}, Z_t)$. The optimal interest rate rule is the one with the G's of (6) chosen so as to minimize loss L subject to (11).[3]

To show that (1)–(5), (8), and (9) yield versions of (10) and (11) that give rise to well-defined, nontrivial dynamic problems, it is enough to examine the one-period reduced forms for y_t and p_t.

With the money supply as the monetary instrument, we solve (1)–(3) for y, r, and p and get as a reduced form for p_t

$$ p_t = J_0({}_tp^*_{t-1}) + J_1({}_{t+1}p^*_{t-1}) + J_2 m_t + X_t, \tag{12} $$

---

[2] See Chow (1970).

[3] Chow (1970) describes how optimal rules of the form (6) or (7) are found for a system like (10) or (11).

where X_t is a linear function (involving the parameters of [1]–[3]) of k_{t-1}, Z_t, and the u_{it}'s and where

$$J_0 = [a_2(1 + b_2c_2^{-1}) + b_2]/[a_2(1 + b_2c_2^{-1}) + b_2c_2^{-1}] < 1,$$
$$J_1 = (1 - J_0)/(1 - c_2^{-1}),$$
$$J_2 = -c_2^{-1}J_1.$$

Substitution of p_t from equation (12) into equation (1) gives the one-period reduced form for y_t. Taking E_{t-1} of p_t and y_t from these reduced forms and eliminating m_t gives the set of pairs $(E_{t-1}y_t, E_{t-1}p_t)$ attainable by choice of m_t. The set is a line whose slope is neither infinity nor zero. Its position, obviously, depends on lagged values of p, via the p^* variables, and on lagged values of other endogenous variables, the distributions of which depend on lagged values of m. In other words, the choice for the deterministic part of m_t has effects in future periods, which is what we mean when we say that (10) gives rise to a nontrivial dynamic problem.

With the interest rate as the monetary instrument, equation (2) is the one-period reduced form for y_t while that for p_t is obtained by substituting the solution for y_t into equation (1) and solving for p_t. The solution for p_t is

$$\begin{aligned} a_2p_t = {} & (a_2 + b_2)_tp^*_{t-1} - b_2({}_{t+1}p^*_{t-1}) + b_2r_t \\ & + (b_1 - a_1)k_{t-1} + b_3Z_t - u_{1t} + u_{2t}. \end{aligned} \tag{13}$$

Again, if we take E_{t-1} of equation (2) and equation (13) and eliminate r_t, we find the set of pairs $(E_{t-1}y_t, E_{t-1}p_t)$ attainable by choice of r_t. That set again depends on lagged values of p, which shows that (11) also gives rise to a nontrivial dynamic problem.

The monetary authority is supposed to solve each of the two dynamic problems, minimizing loss first under an m rule and then under an r rule. Which policy is superior depends on which delivers the smaller loss, which in turn depends on all of the parameters of the model, including the covariance matrix of the disturbances. Which rule is superior is therefore an empirical matter, an outcome which is completely consistent with Poole's analysis.

4. The Rational-Expectations Version under a Money Supply Rule

Here we impose the requirement that the public's expectations be rational by requiring that

$${}_{t+i}p^*_{t-j} = E_{t-j}p_{t+i}, \tag{14}$$

where $E_{t-j}p_{t+i}$ is the mathematical expectation of p_{t+i} calculated using the model (i.e., the probability distribution of p_{t+i}) and all information assumed to be available as of the end of period $t - j$. The available information is assumed to consist of data on current and past values of all endogenous and exogenous variables observed as of the end of period $t - j$, that is, θ^*_{t-j}.

To begin, we again solve the system (1)–(3) for y, r, and p given m. With expectations given by (14), what is now a pseudo-reduced-form equation for p is

$$p_t = J_0 E_{t-1} p_t + J_1 E_{t-1} p_{t+1} + J_2 m_t + X_t. \tag{15}$$

Computing $E_{t-1} p_t$ from (15) and subtracting the result from (15) we get

$$\begin{aligned} p_t - E_{t-1} p_t &= J_2(m_t - E_{t-1} m_t) + X_t - E_{t-1} X_t \\ &= X_t - E_{t-1} X_t, \end{aligned} \tag{16}$$

where the last equality follows from the assumption that a deterministic rule of the form (7) is being followed. But since $X_t - E_{t-1}X_t$ is a linear combination of the innovations in the exogenous processes, it follows that $p_t - E_{t-1}p_t$ is an exogenous process, unaffected by the rule chosen for determining the money supply.

Using (14) and (16), we can write equation (1) as

$$y_t = a_1 k_{t-1} + a_2(X_t - E_{t-1} X_t) + u_{1t}. \tag{17}$$

If we substitute the right-hand side for y_t in equation (2), we can obtain the real interest rate as a function of k_{t-1} and exogenous processes. Substituting that function into equation (4), we get a difference equation in k driven by exogenous processes. This proves that k is an exogenous process, which by (17) implies that y is an exogenous process, that is, has a distribution independent of the deterministic rule for the money supply. So we have proved assertion (*a*) above: the distribution of output does not depend on the parameters of the feedback rule for the money supply.

To prove assertion (*b*), we write the tth term of the loss function L as

$$L_t = E_0[E_{t-1}(K_2 p_t + K_{22} p_t^2 + K_1 y_t + K_{11} y_t^2)],$$

where the insertion of E_{t-1} is valid for $t > 0$. Using $E(x^2) = E[(x - Ex)^2] + (Ex)^2$, we have

$$L_t = E_0[K_{0t} + K_2 E_{t-1} p_t + K_{22}(E_{t-1} p_t)^2],$$

where

$$K_{0t} = E_{t-1}[K_{22}(p_t - E_{t-1} p_t)^2 + K_1 y_t + K_{11} y_t^2]$$

and where, given the exogeneity of y_t and $p_t - E_{t-1}p_t$ proved above, K_0 is an exogenous process. Moreover, it is possible, as we shall show below, to find a rule for m that implies choosing $E_{t-1}p_t$ to minimize

$$K_{0t} + K_2 E_{t-1} p_t + K_{22}(E_{t-1} p_t)^2.$$

And, because K_{0t} is unaffected by settings for the money supply at any other time, a rule which minimizes L_t also minimizes L.

To show that there exists such a rule for m, we must solve the model. Again, we take $E_{t-1}p_t$ in (15) and write the result as

$$(1 - J_0)E_{t-1}p_t = J_1E_{t-1}p_{t+1} + J_2E_{t-1}m_t + E_{t-1}X_t. \qquad (18)$$

Since this holds for all t, it follows that

$$(1 - J_0)E_{t-1}p_{t+j} = J_1E_{t-1}p_{t+j+1} + J_2E_{t-1}m_{t+j} + E_{t-1}X_{t+j}. \qquad (19)$$

By repeated substitution from (19) into (18), we obtain

$$(1 - J_0)E_{t-1}p_t = \sum_{j=0}^{n} [J_1/(1 - J_0)]^j(E_{t-1}X_{t+j} + J_2E_{t-1}m_{t+j}) + [J_1/(1 - J_0)]^{n+1}E_{t-1}p_{t+n+1}, \qquad (20)$$

where

$$0 < J_1/(1 - J_0) = 1/(1 - c_2^{-1}) < 1.$$

We assume that the limit as $n \to \infty$ of the second term on the right-hand side of (20) is zero, which is a terminal condition that has the effect of ruling out "speculative bubbles." Then, from (20),

$$(1 - J_0)E_{t-1}p_t = \sum_{j=0}^{\infty} [J_1/(1 - J_0)]^jE_{t-1}(X_{t+j} + J_2m_{t+j}). \qquad (21)$$

Since this holds for all t, we may replace t by $t + 1$ and compute E_{t-1} of the result to get

$$(1 - J_0)E_{t-1}p_{t+1} = \sum_{j=0}^{\infty} [J_1/(1 - J_0)]^j \times E_{t-1}(X_{t+j+1} + J_2m_{t+j+1}). \qquad (22)$$

For an arbitrary money supply rule of the form (7), substituting (21) and (22) into (15) gives the solution for p_t; substituting (21) and (22) into (2) gives the solution for r_t. This assumes that the rule is not such as to imply too explosive a process for $X_{t+j} + J_2m_{t+j}$.[4]

---

[4] A workable "reduced form" for p_t can be obtained by substituting (20) into (15) and then by using (5) and (7) to replace $E_{t-1}m_{t+j}$ and $E_{t-1}X_{t+j}$ with the linear functions of past variables that they equal. These linear functions are easily calculated from the feedback rule for m_t and the autoregressions governing components of X_t. While the resulting "reduced form" for p_t formally resembles the corresponding equation in the system with "adaptive" expectations, there is a crucial difference. Now changes in the parameters of the feedback rule for m_t produce changes in the parameters of the reduced form for p_t. This feature of the system is what renders Poole's results inapplicable. For an explicit illustration of the dependence of the reduced-form parameters on the form of the policy rule, see Sargent and Wallace (1973, pp. 332–33).

To find the optimal money supply rule, multiply (22) by $J_1/(1 - J_0)$ and subtract the result from (21) to get

$$(1 - J_0)E_{t-1}p_t - J_1E_{t-1}p_{t+1} = E_{t-1}X_t + J_2m_t. \tag{23}$$

The value of $E_{t-1}p_t$ that minimizes L_t for all t is

$$E_{t-1}p_t = -K_2/2K_{22}, \tag{24}$$

so that

$$E_{t-1}p_{t+1} = -K_2/2K_{22}. \tag{25}$$

The optimal rule for the money supply is obtained by substituting (24) and (25) into (23). The resulting expression for m_t is a feedback rule of the form (7).

Thus, in the rational-expectations version of our model, the choice among deterministic rules for the money supply is a trivial problem. One argument of the loss function, y, is unaffected by the rule, so the problem becomes the simplest kind of one target–one instrument problem. Moreover, a definite rule emerges only because we have assumed in specifying L that there is a target value for the price level. If, instead, loss were made dependent only on the variance of the price level, then one deterministic rule would be as good as any other.

The reason that the distribution of real output is independent of the systematic money supply rule can be summarized as follows. In order for the monetary authority to induce fluctuations in real output, it must induce unexpected movements in the price level by virtue of the aggregate supply curve (1). But by virtue of the assumption that expectations about the price level are rational, the unexpected part of price movements is independent of the systematic part of the money supply, as long as the authority and the public share the same information. There is no systematic rule that the authority can follow that permits it to affect the unexpected part of the price level. Of course, the authority could add an unpredictable random term to the systematic part of the money supply, so that (7) would be amended to become

$$m_t = H\theta^*_{t-1} + \psi_t, \tag{7'}$$

where ψ_t is a random variable obeying $E\psi_t \mid \theta^*_{t-1} = 0$. Then the distribution of unexpected price movements and of real output will depend on the distribution of ψ_t. But clearly, there is no way the authority can base a countercyclical policy on this particular nonneutrality, since there is no way the authority can regularly choose ψ_t in response to the state of economic affairs in order to offset other disturbances in the system.

This follows since ψ_t is that part of the money supply obeying $E\psi_t \mid \theta^*_{t-1} = 0$. Furthermore, in our model it is optimal to set $\psi_t = 0$ for all t.

5. The Rational-Expectations Version under an Interest Rate Rule

Above we showed that a certain terminal condition implied the existence of a unique equilibrium price level for the rational-expectations version under a money supply rule that is not too explosive. That analysis took as a starting point the difference equation (18). With the interest rate determined by the feedback rule (6), a seemingly analogous difference equation is obtained by imposing rationality, equation (14), in (13) and taking E_{t-1} of the result

$$\begin{aligned} 0 = {} & b_2(E_{t-1}p_t - E_{t-1}p_{t+1}) + b_2 r_t + (b_1 - a_1)k_{t-1} \\ & + b_3 E_{t-1}(Z_t - u_{1t} + u_{2t}). \end{aligned} \tag{26}$$

If we solve (26) by recursion, proceeding as we did in deriving (20) from (18), we find

$$\begin{aligned} E_{t-1}p_t = {} & -\sum_{j=0}^{n} E_{t-1}\{r_{t+j} + [(b_1 - a_1)/b_2]k_{t+j-1} + (b_3/b_2) \\ & \times (Z_{t+j} - u_{1t+j} + u_{2t+j})\} + E_{t-1}p_{t+n+1}. \end{aligned} \tag{27}$$

To obtain a particular solution for $E_{t-1}p_t$ from (27) requires imposing a terminal condition in the form of taking as exogenous a value of $E_{t-1}p_{t+j}$ for some $j \geq 0$. This is obviously a very much stronger terminal condition than we had to impose on (20), a consequence of (26) being a nonconvergent difference equation. Thus, when the interest rate is pegged, the model cannot determine a path of expected prices $E_{t-1}p_{t+j}$, $j = 0, 1, \ldots$, and by implication cannot determine the price level p_t. Neither can it determine the money supply.

The economics behind the underdetermined expected price level is pretty obvious. Under the interest rate rule (6), the public correctly expects that the monetary authority will accommodate whatever quantity of money is demanded at the pegged interest rate. The public therefore expects that, *ceteris paribus*, any increase in p_t will be met by an equal increase in m_t. But that means that one $E_{t-1}p_t$ is as good as any other from the point of view of being rational. There is nothing to anchor the expected price level. And this is not simply a matter of choosing the "wrong" level or rule for the interest rate. There is no interest rate rule that is associated with a determinate price level.

At least since the time of Wicksell it has been known that, in the context of a static analysis of a full-employment model with wages and prices that are flexible instantaneously, it can happen that the price level is indeterminate if the monetary authority pegs the interest rate.[5] In such a static analysis, the indeterminacy of the price level depends critically on output and employment being exogenous with respect to shocks to aggregate demand or portfolio balance; that is, the Phillips curve must be vertical. In our model, however, the Phillips curve is not vertical, but Wicksell's indeterminacy still arises.

6. An Information Advantage for the Monetary Authority

Here we shall examine some consequences of the monetary authority having more information than the public. We shall first show that if the monetary authority follows the money supply rule that is optimal if there is no information discrepancy, then the loss attained is the same as attained when there is no information discrepancy. Then we consider whether that rule is optimal given an information discrepancy.

We shall write E_{t-1} for the expectation conditional on what the monetary authority knows at the end of period $t - 1$ and $E_{\theta,t-1}$ for the expectation conditional on what the public knows at the end of period $t - 1$, where θ is a subset of what the monetary authority knows. Then in place of (14) we impose

$$ {}_{t+i}p^{*}_{t-j} = E_{\theta,t-j}p_{t+i}, \tag{28}$$

so that in place of (15) we have

$$p_t = J_0E_{\theta,t-1}p_t + J_1E_{\theta,t-1}p_{t+1} + J_2m_t + X_t. \tag{29}$$

Then, taking $E_{\theta,t-1}$ of p_t and subtracting from p_t, we have

$$p_t - E_{\theta,t-1}p_t = J_2(m_t - E_{\theta,t-1}m_t) + (X_t - E_{\theta,t-1}X_t). \tag{30}$$

The rule that we found to be optimal without an information discrepancy is

$$J_2m_t = -(K_2/2K_{22})(1 - J_0 - J_1) - E_{t-1}X_t. \tag{31}$$

From this it follows that

$$J_2(m_t - E_{\theta,t-1}m_t) = -E_{t-1}X_t + E_{\theta,t-1}X_t. \tag{32}$$

Substituting into (30), we have

$$p_t - E_{\theta,t-1}p_t = X_t - E_{t-1}X_t. \tag{33}$$

---

[5] See Olivera (1970). In both our model and the standard static model, the aggregate demand schedule must exclude any components of real wealth that vary with the price level if Wicksell's indeterminacy is to arise. For example, if the anticipated rate of capital gains on real (outside) money balances is included in the aggregate demand schedule, the price level is determinate with a pegged interest rate. However, such a system has peculiar stability characteristics, since stability hinges on the sign of the expected rate of inflation.

Upon substituting from (33) into equation (1) of the structure, we get equation (17). And if we substitute for y_t in equation (2) the right-hand side of (17), we obtain $[r_t - E_{\theta,t-1}(p_{t+1} - p_t)]$ as a function of k_{t-1} and the exogenous processes, the same function that we previously found for $r_t - E_{t-1}(p_{t+1} - p_t)$. Then, substituting this function into equation (5) of the structure, we get the same first-order difference equation in k as we had without an information discrepancy. This proves that under the rule given by (31), the distribution of k does not depend on the information discrepancy. It follows then from equation (17) that the same is true for y.

To find the distribution of p_t, we proceed to solve the difference equation

$$(1 - J_0)E_{\theta,t-1}p_t = J_1E_{\theta,t-1}p_{t+1} + J_2E_{\theta,t-1}m_t + E_{\theta,t-1}X_t, \qquad (34)$$

which is obtained by taking $E_{\theta,t-1}$ of (29). Then, proceeding as we did for (18), we obtain expressions exactly like (21) and (22) except that in place of E_{t-1} on the left and right we have $E_{\theta,t-1}$.

But from (31)

$$X_{t+j} + J_2m_{t+j} = -(K_2/2K_{22})(1 - J_0 - J_1) + X_{t+j} - E_{t-1}X_{t+j},$$

so for $j > 0$

$$\begin{aligned} E_{\theta,t-1}(X_{t+j} + J_2m_{t+j}) &= -(K_2/2K_{22})(1 - J_0 - J_1) \\ &= E_{t-1}(X_{t+j} + J_2m_{t+j}). \end{aligned}$$

Thus, use of the rule given by (31) implies $E_{\theta,t-1}p_{t+j} = E_{t-1}p_{t+j}$, $j = 0, 1$, which by (29) implies that under the rule given by (31) the distribution of p does not depend on θ, that is, does not depend on the information discrepancy. It follows that the loss attained under the rule given by (31) does not depend on the information discrepancy.

This shows that the monetary authority can do as well given an information discrepancy as it can do if there is none. But can it do better? Can it, as it were, take advantage of the presence of an information discrepancy? We are not sure. But within our structure, the answer seems to be that it can take advantage of a discrepancy, although necessarily in a limited and rather subtle way.

To indicate why, let us focus first on how the distribution of y depends on the rule for m. Under present assumptions, equation (1) of the structure is

$$y_t = a_1k_{1-t} + a_2(p_t - E_{\theta,t-1}p_t) + u_{1t}. \qquad (35)$$

It follows that as of the end of $t - 1$, $E_{\theta,t-1}y_t$ is unaffected by the choice of m_t, since

$$E_{\theta,t-1}y_t = a_1E_{\theta,t-1}k_{t-1} + E_{\theta,t-1}u_{1t}. \qquad (36)$$

To find the variance of y_t, we subtract (36) from (35) and obtain $\tilde{y}_t = a_1\tilde{k}_{t-1} + a_2\tilde{p}_t + \tilde{u}_{1t}$, where $\tilde{x}_t \equiv x_t - E_{\theta,t-1}x_t$. The variance of

y_t around $E_{\theta,t-1}y_t$ is, therefore,

$$E_{\theta,t-1}(\tilde{y}_t^2) = E_{\theta,t-1}[(a_1\hat{k}_{t-1} + \hat{u}_{1t})\hat{p}_t] + E_{\theta,t-1}(\hat{p}_t^2) \\ + \text{ other terms,} \tag{37}$$

where $\hat{x}_t \equiv E_{t-1}x_t - E_{\theta,t-1}x_t$ and where the omitted terms are unaffected by the setting for the deterministic part of m_t. Thus, setting m_t according to (31) (i.e., setting $\hat{p}_t = 0$) minimizes $E_{\theta,t-1}(\tilde{y}_t^2)$ only if the first term on the right-hand side of (37) cannot be made negative. That term can be made negative by a rule different from (31) if $a_1\hat{k}_{t-1} + \hat{u}_{1t} \neq 0$, that is, if the monetary authority knows more about either the k_{t-1} or the u_1 process than does the public. Of course, to take advantage of this information discrepancy, the monetary authority must know precisely how the public's information differs from its own.

Similar conclusions hold for the distribution of k_t. The expectation $E_{\theta,t-1}k_t$ is unaffected by the setting for m_t, but, in general, the variance $E_{\theta,t-1}(\tilde{k}_t^2)$ depends on it and is not minimized by use of the rule given by (31).[6] And since the setting for m_t affects the distribution of (y_{t+j}, p_{t+j}) for $j > 0$ by way of its effect on the distribution of k_t, this means that, given an information discrepancy, our structure gives rise to a nontrivial dynamic problem.

But this should not be taken to mean that we are back in the setting produced by the assumption that expectations are formed on the basis of fixed autoregressive schemes. The information discrepancy assumption does not produce any simple tradeoff between the means of output and the price level. The fact that $E_{\theta,t-1}y_t$ and $E_{\theta,t-1}k_t$ are unaffected by m_t is very limiting if θ contains, say, as little as $(1, p_{t-1}, y_{t-1})$. Second, to exploit the information discrepancy, the monetary authority must know what it is. To assume that it does seems farfetched. Indeed, we suspect that estimating the discrepancy is a very subtle and perhaps intractable econometric problem.

For these reasons, we think some comfort can be taken from the first result established in this section. Use of the rule given by (31) is optimal if the public is as well informed as the monetary authority. The loss attained under that rule does not depend on how well informed the public is, and implementation of the rule does not required knowledge of how well informed the public is.

This does not, of course, deny that there is a gain from learning more about the exogenous processes. Settings for the money supply under the rule given by (31) depend on what the monetary authority knows. Operating under that rule, loss is smaller the more the monetary authority knows about the exogenous processes.

---

[6] The reader may verify this by finding k_t as a function of k_{t-1} and $p_t - E_{\theta,t-1}p_t$ using (35) and equations (2) and (5) of the structure.

7. Concluding Remarks

Given that our conclusions are derived from an ad hoc model, should they be taken seriously? In one sense, they should not be. Because of their ad hoc nature, neither the structure set out in section 1 nor the loss function of section 2 should be accepted as providing a suitable context within which to study macroeconomic policy. Nevertheless, some aspects of our model cannot be dismissed so easily. First, the hypothesis that expectations are rational must be taken seriously, if only because its alternatives, for example, various fixed-weight autoregressive models, are subject to so many objections. Second, the aggregate supply hypothesis is one that has some microeconomic foundations,[7] and it has proved difficult to dispose of empirically.[8] It is precisely these two aspects of our model—rational expectations in conjunction with Lucas's aggregate supply hypothesis—that account for most of our results. We believe that the results concerning systematic countercyclical macroeconomic policy are fairly robust to alterations of other features of the model, such as the aggregate demand schedule and the portfolio balance condition. In particular, the dramatically different implications associated with assuming rational expectations, on the one hand, or fixed autoregressive expectations, on the other hand, will survive such alterations.

References

Chow, Gregory C. "Optimal Stochastic Control of Linear Economic Systems." *J. Money, Credit and Banking* 2 (August 1970): 291–302.

Lucas, Robert E., Jr. "Some International Evidence on Output-Inflation Tradeoffs." *A.E.R.* 63 (June 1973): 326–34.

Muth, John F. "Rational Expectations and the Theory of Price Movements." *Econometrica* 29 (July 1961): 315–35.

Olivera, Julio H. "On Passive Money." *J.P.E.* 78, no. 4, suppl. (July/August 1970): 805–14.

Poole, William. "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model." *Q.J.E.* 84 (May 1970): 197–216.

Sargent, Thomas. *"Rational" Expectations, the Real Rate of Interest, and the "Natural" Rate of Unemployment.* Brookings Papers on Economic Activity, no. 2. Washington: Brookings Inst., 1973.

Sargent, Thomas, and Wallace, Neil. "Rational Expectations and the Dynamics of Hyperinflation." *Internat. Econ. Rev.* 14 (June 1973): 328–50.

---

[7] For example, see Lucas (1973).

[8] Tests of the aggregate supply hypothesis are reported by Lucas (1973) and Sargent (1973).

The Relevance of the Household Production Function and Its Implications for the Allocation of Time

---

Robert A. Pollak and Michael L. Wachter
University of Pennsylvania

This paper provides a critique of the household production function approach and its application to the allocation of time. It is argued that many applications of the model, especially those making use of implicit "commodity prices," require that the household's technology exhibit constant returns and no joint production; otherwise, implicit commodity prices depend on the household's preferences as well as on its technology and the prices of market goods. Furthermore, joint production is pervasive in situations involving the allocation of time. In situations where household production theory does not provide a satisfactory framework for analysis, the paper suggests alternative approaches.

1. Introduction

Recent work by Becker, Lancaster, and others provides a new foundation for the theory of household behavior. According to the new view, the household purchases "goods" on the market and combines them with time in a "household production function" to produce "commodities." These commodities, rather than the goods, are the arguments of the household's utility function; market goods and time are not desired for their own sake, but only as inputs into the production of "commodities."[1]

The authors are professor and associate professor of economics at the University of Pennsylvania. The research was supported by grants from the National Science Foundation and the National Institute of Child Health and Human Development. We wish to thank James J. Heckman, Michael Landsberger, John Muellbauer, Marc Nerlove, Stephen A. Ross, and the anonymous referees for helpful comments.

[1] The production function approach has been applied to a wide range of problems. Becker (1965) suggests that it be used to study such commodities as "seeing a play," which depends on the input of actors, script, theater, and the playgoer's time, and "sleeping," which depends on the input of a bed, house, and time. He has also used it (Becker 1960) to study fertility, introducing a commodity, the "quality of children," which is produced by inputs of goods and parental time. In Lancaster's model (1966*a*, 1966*b*, 1971) goods possess "characteristics"; these characteristics, which we can identify with Becker's commodities, are the arguments of the household's utility function. Each unit of a good, for example, a "glass of orange juice," produces a vector of characteristics such as calories and vitamin C.

[*Journal of Political Economy*, 1975, vol. 83, no. 2]
© 1975 by The University of Chicago. All rights reserved.

The new approach is oriented toward the study of nonmarket commodities and the associated allocation of household time—topics which have been neglected in the traditional theory of consumer behavior. Furthermore, stress is placed on the role of the household's technology as well as its tastes as a determinant of behavior. Typically, the analysis focuses on the demand for commodities as functions of "commodity prices," which, in turn, depend on goods prices and the household's technology.

In this paper we argue that, except in very special cases, the new theory does not provide a satisfactory model of the demand for commodities and the allocation of time as functions of "commodity prices." We identify the cases in which the new theory can do so, and propose alternative approaches for those in which it does not. We first analyze the conditions under which the household production function theory allows one to study nonmarket commodities as functions of commodity prices. We argue that this application of the household production function approach requires strong assumptions about the structure of the household's technology, in particular, constant returns to scale and the absence of joint production. If these assumptions are not satisfied, commodity prices depend on the household's consumption pattern. Hence, price differences among households reflect differences in tastes as well as technology. Since commodity prices are not determined solely by goods prices and technology, but also reflect the consumption pattern chosen by the household, it is misleading to treat the demand for commodities as a function of "commodity prices." Commodities which involve time as an input are a particular problem, since their production often involves joint production. Jointness is pervasive because time spent in many production activities is a *direct* source of utility as well as an input into a commodity. We suggest that the household production function model, with its concentration on commodity prices, provides a satisfactory account of the allocation of time only for households which are indifferent among alternative uses of their time. When joint production and/or nonconstant returns to scale are present, we propose that commodity prices be avoided and that the demand for commodities be viewed as a function of goods prices, the wage rate, and nonlabor income.

An additional problem occurs when the household production function model is applied to variables which may be interpreted as "utilities" (numbers representing preference orderings) rather than "commodities" (the outputs of production processes). In cases of this type, the production function approach loses its unique identity and cannot be distinguished from a variety of hypotheses about the structure of the household's preferences. Activities which generate "utilities" can be studied, but the analysis must focus on the household's allocation of goods and time among these activities rather than their "output." The resulting analytical framework is closer to traditional demand theory than to the household production function model.

We proceed as follows: in Section 2 the basic household production function model is presented and its implications for the demand for goods and the demand for commodities are examined. Our discussion in Section 2 assumes that the household's technology exhibits constant returns to scale and no joint production. The consequences of relaxing these assumptions are discussed in Section 3. In Section 4 we consider the role of time in the household production function. In Section 5 alternative approaches to studying nonmarket activities and the allocation of time are suggested.

2. Household Production Functions

Suppose that there are n goods, denoted by $X = (x_1, \ldots, x_n)$, and m commodities, denoted by $Z = (z_1, \ldots, z_m)$. The goods are inputs into the production of commodities. We ignore until Section 4 the presence of household time as an input along with market goods. In this section we assume that each commodity is produced by a separate constant-returns-to-scale production function. This assumption rules out joint production and enables us to identify the inputs used to produce each commodity. Let y_{ri} denote the quantity of the ith good used as an input in the production of the rth commodity, and let Y_r denote the corresponding vector: $Y_r = (y_{r1}, \ldots, y_{rn})$.[2] Denote the production function for the rth commodity by $f^r: z_r = f^r(Y_r) = f^r(y_{r1}, \ldots, y_{rn})$. Constant returns to scale implies $f^r(\lambda Y_r) = \lambda f^r(Y_r)$. Since every good need not be used in the production of every commodity, some of the y_{ri}'s may be zero. The goods vector X is related to the input vectors by

$$X = \sum_{s=1}^{m} Y_s.$$

In the household production function model, the household's preferences are represented by a utility function $U(Z)$ defined over the commodity space. Goods are not desired for their own sake, but only because they are inputs for the production of commodities. In the household production function model the household faces two types of constraints on its consumption opportunities: the budget constraint and the limitation imposed by its technology. Hence, the demand for goods and the demand for commodities both depend on goods prices, the household's income, its tastes, and its technology.

We define the cost function $C(P, Z)$ as the cost of the least expensive collection of goods capable of producing the commodity vector Z when goods prices are P. Formally

---

[2] Although Becker often uses fixed-coefficient production functions as an expositional device, we shall not, since fixed coefficients are not an integral part of the household production function model.

$$C(P, Z) = \min \sum_{k=1}^{n} p_k x_k$$

subject to $(X, Z) \in \Omega$, where Ω is the "production set" whose elements are technically feasible input-output vectors. We hereafter denote the feasible set of commodity vectors producible from the goods vector X by $S(X)$.

In the household production function model primary stress is placed on the demand functions for commodities rather than the demand functions for goods. Following this literature, we define implicit commodity prices $(\pi_1, \ldots, \pi_m) = \pi$ as the marginal costs of producing commodities:

$$\pi_r(P, Z) = \frac{\partial C(P, Z)}{\partial Z_r}.$$

Under certain conditions these implicit commodity prices are independent of the z's, so $\pi_r(P, Z) = \pi_r(P)$.[3] When they are, the budget constraint in the commodity space can be written as

$$\sum_{s=1}^{m} \pi_s(P) z_s = \mu.$$

Maximizing $U(Z)$ subject to this budget constraint yields commodity demand functions $z_r = g^r(\pi, \mu)$, or, in vector form $Z = g(\pi, \mu)$. These demand functions exhibit all of the properties of traditional demand theory; they are homogeneous of degree zero in (π, μ) and satisfy the budget constraint and the Slutsky sign and symmetry conditions. Since the household production function model places no special restrictions on the utility function $U(Z)$, it implies no additional restrictions on the commodity demand functions.

If we are to preserve the formal analogy with traditional demand theory, it is crucial that the commodity prices be independent of the commodity bundle which the household chooses. The advantage of preserving this formal correspondence is that by doing so we have access to the body of established theoretical results of traditional demand theory. If implicit commodity prices depend on the commodity bundle consumed, then the commodity demand relations correspond to those in a model in which consumers are monopsonists or are offered tie-in sales, and there are virtually no substantive results for these cases. In Section 3, we discuss the case in which commodity prices violate this condition and depend on the z's. We now identify a case in which commodity prices are independent of the commodity bundle consumed.

---

[3] Much of this section and the next is devoted to investigating the conditions under which implicit commodity prices are independent of the Z's.

Theorem: **If the household's technology exhibits constant returns to scale and no joint production, then the budget constraint in the commodity space is of the form**

$$\sum_{s=1}^{m} \pi_s z_s = \mu,$$

where the π's depend on goods prices and the household's technology, but not on the commodity bundle consumed.

Proof: With no joint production, we can write the cost function $C(P, Z)$ as

$$C(P, Z) = \sum_{s=1}^{m} C^s(P, z_s),$$

where $C^r(P, z_r)$ is the cost function associated with the production function $z_r = f^r(Y_r)$. Constant returns to scale implies

$$C^r(P, z_r) = C^r(P, 1)z_r.$$

Hence the cost function becomes

$$C(P, Z) = \sum_{s=1}^{m} C^s(P, 1)z_s$$

and

$$\pi_r = C^r(P, 1). \qquad \text{Q.E.D.}$$

The π's reflect both the goods prices which the household faces and the household's technology, but they are independent of the household's tastes. They represent the marginal cost of each commodity to the household, and, under our assumptions of constant returns to scale and no joint production, they are independent of the quantity of the commodity in question as well as the quantities of all other commodities consumed by the household.

Although the household production function literature places primary stress on the demand for commodities, the implications of the model for the demand for goods also must be developed. To find the demand functions for goods, the most direct approach is to determine the utility function for goods implied by the household's technology and its utility function for commodities and then to maximize this goods utility function subject to the appropriate budget constraint. We begin by constructing the utility function for goods, $V(X)$. The household's technology allows us to associate a set of feasible commodity vectors, $S(X)$, with every goods vector X. Commodity vectors in $S(X)$ are producible from the input vector X; those not in $S(X)$ are not producible. The utility function, $U(Z)$, is used to select a best commodity vector Z^* from this feasible set. Hence, one can define the utility function over the goods space, $V(X)$, by associating with the goods vector X the utilitv of the commodity vector Z^*; that is, $V(X) = \max U(Z)$ subject to $Z \in S(X)$.[4] Finally, to find the goods demand functions, maximize $V(X)$ subject to the budget constraint

$$\sum_{k=1}^{n} p_k x_k = \mu,$$

where μ denotes total expenditure or "income." This yields the "goods demand functions," $x_i = h^i(P, \mu)$, or, in vector form, $X = h(P, \mu)$, where P is the vector of goods prices. That is, the household's technology and its utility function over the commodity space imply a utility function over the good space; maximizing that utility function subject to the budget constraint yields demand functions for goods. It should be emphasized that $V(X)$, the translation of the utility function into the goods space, reflects both tastes and technology. Traditional demand theory treats $V(X)$ as the household's utility function and is thus guilty of confounding tastes and technology, rather than maintaining a separation between them. A consequence of this, as the household production function literature points out, is that changes in demand which are clearly attributable to changes in technology must formally be described as changes in tastes.

The household production function model emphasizes several key points which are overlooked in the traditional theory of the household. First, it develops a framework for analyzing the allocation of goods (and time) within the household. Second, it stresses the role of the household's technology as well as its tastes as a determinant of behavior. This opens the possibility of attributing variations in household behavior to changes in technology as well as to changes in goods prices or income operating through the budget constraint or to changes in tastes. Third, the household production function model focuses our attention on the outputs of the household technology, the commodities, while the traditional theory ignores commodities and considers only market goods.

From the standpoint of the household, the budget constraint and the technological constraints are coequal limitations of its consumption possibilities. But from the standpoint of the economist investigating household behavior, there is a crucial difference between them: the household's budget constraint is directly observable, while the household's technology must be inferred from its behavior. To construct commodity prices which can be utilized as arguments of these demand functions, the parameters of the technology must be known. A procedure for implementing the household production function model involves two stages. The first stage includes estimation of the household's technology and construction of commodity prices from goods prices using the estimated parameters of the technology. These commodity prices serve as arguments of the commodity demand functions, $Z = g(\pi, \mu)$; the estimation of these commodity demand functions constitutes the second stage.

---

[4] Mild regularity conditions on the production functions, together with the assumption of constant returns to scale, imply that the production set Ω is convex. If the production set Ω is convex and the commodity utility function $U(Z)$ is quasi-concave, then the implied goods utility function $V(Z)$ is quasi-concave. The proof is straightforward. Notice that the convexity assumption applies to the production set Ω and that its convexity implies the convexity of $S(X)$.

There are three approaches to investigating the household's technology. One is direct estimation of the production functions using data on imputs of goods and outputs of commodities. The second is the calculation of the total or unit cost functions; these yield estimates of commodity prices at various configurations of goods prices. The third is to estimate the "constant output factor demand functions," which show the demand for goods as a function of goods prices and commodity outputs. Since these demand functions are the partial derivatives of the cost functions with respect to goods prices, and since the cost functions are homogeneous of degree 1 in goods prices, the cost functions can be constructed from them by Euler's theorem.

Once commodity prices have been calculated, the household production function approach can use them as independent variables in the analysis of the demand for commodities. Because the commodity prices depend on the household's technology as well as on goods prices, they can be used to express the fundamental insight of the household production function model, namely, that the household is constrained by technology as well as by its limited budget, and that changes in technology provide a possible explanation of changes in household behavior.

3. Complications

In this section we consider the consequences of joint production and nonconstant returns to scale for the household production function model. We show that either one implies that commodity prices depend on the commodity bundle consumed and argue that this impares the analytical usefulness of commodity prices as a tool for analyzing the demand for commodities.

Assume that the household's technology can be represented by a generalized production function $z_1 = T(z_2, \ldots, z_m, x_1, \ldots, x_n)$. Unless explicitly stated to the contrary, constant returns to scale and/or the absence of joint production are *not* assumed.

Even with joint production and nonconstant returns to scale, there is no difficulty with the demand functions for goods. To find them, we define the implied goods utility function, $V(X)$, by $V(X) = \max U(Z)$ subject to $z_1 = T(z_2, \ldots, z_m, x_1, \ldots, x_n)$. Maximizing $V(X)$ subject to the budget constraint $\sum p_k x_k = \mu$ yields the goods demand functions

$X = h(P, \mu)$. If the utility function $V(X)$ is quasi-concave, these demand functions exhibit all the properties of demand functions of traditional demand theory, and joint production and nonconstant returns to scale have no impact on the general properties of the goods demand functions. But the quasi-concavity of $V(X)$ does not follow from our assumptions about $U(Z)$ unless we place further restrictions on the technology.[5]

To determine the commodity demand functions, we translate the budget constraint from the goods space into the commodity space. The cost function $C(P, Z)$ is defined as the minimum cost of producing the commodity bundle Z at goods prices P; that is, $C(P, Z) = \min \sum p_k x_k$, subject to $z_1 = T(z_2, \ldots, z_m, x_1, \ldots, x_n)$. The household maximizes $U(Z)$ subject to the translated budget constraint $C(P, Z) = \mu$. The first-order conditions imply

$$\frac{\partial U(Z)}{\partial z_r} = \lambda \frac{\partial C(P, Z)}{\partial z_r} \qquad r = 1, \ldots, m,$$

and we again interpret the partial derivatives of the cost function as implicit commodity prices; that is, the marginal cost of each commodity is its implicit or shadow price. Except in special cases, these implicit prices depend on the household's technology, the goods prices, *and the commodity bundle consumed by the household.*

We define the "implicit income" of the household consuming the commodity bundle Z at goods prices, P, $I(P, Z)$, by

$$I(P, Z) = \sum_{s=1}^{m} z_s \frac{\partial C(P, Z)}{\partial z_s} .$$

That is, $I(P, Z)$ is the cost of the commodity bundle Z evaluated at the implicit commodity prices. If the household's technology exhibits constant returns to scale, then $C(P, Z) = \mu$ implies $I(P, Z) = \mu$ by Euler's theorem. Without constant returns, $I(P, Z) \neq \mu$; the discrepancy depends on the relative importance of increasing and decreasing returns.

When commodity prices depend on the commodity bundle consumed, we do not believe that it is analytically useful to view the demand for commodities as a function of commodity prices and implicit income. Since the π's and I depend on the commodity bundle consumed, they cannot be treated as predetermined parameters from the standpoint of the household. In Section 2 we showed that the implicit commodity prices are independent of the commodity bundle consumed if the household's technology exhibits both constant returns to scale and no joint production. We now show the converse: in the absence of fixed costs, this is the only case in which implicit commodity prices are independent of the z's.

---

[5] The problem is with increasing returns rather than joint production. If the production set Ω is convex, it is easy to show that the quasi-concavity of $U(Z)$ implies the quasi-concavity of $V(X)$. With increasing returns, however, the production set is not convex, and it may be so nonconvex as to yield a goods utility function which is not quasi-concave. For example, suppose $U(z_1, z_2) = z_1^{1/2} + z_2^{1/2}$ and the household's technology is given by the production functions $z_1 = x_1^4$ and $z_2 = x_2^4$. The implied goods utility function is $V(x_1, x_2) = x_1^2 + x_2^2$, which is clearly not quasi-concave.

Theorem: Suppose the household's technology is continuous, and $T(0,\dots,0,0,\dots,0) = 0$. Then implicit commodity prices are independent of the commodity bundle consumed only if the household's technology exhibits both constant returns to scale and no joint production.[6]

Proof: If implicit commodity prices are independent of the commodity bundle consumed, then the cost function can be written as

$$C(P, Z) = C^{o}(P) + \sum_{s=1}^{n} C^{s}(P)z_{s}.$$

Since the technology is continuous and $T(0,\dots,0,0,\dots,0) = 0$, $C(P, 0) = 0$, and, hence, $C^{o}(P) = 0$. To complete the proof, one must show that $C^{r}(P)$ is a unit cost function corresponding to some constant-returns-to-scale production function. A formal proof can be found in Diewert (1972) and Hall (1973). Q.E.D.

Thus far we have focused on the conditions under which *absolute* commodity prices, $\pi_r = [\partial C(P, Z)]/\partial z_r$, are independent of the commodity bundle consumed. It is possible, however, for *relative* commodity prices $(\partial C/\partial z_r)/(\partial C/\partial z_s)$ to be independent of the commodity bundle consumed, even when absolute commodity prices are not.[7] An example of a case in which relative commodity prices are independent of the z's but absolute prices are not is given by the technology

$$\left(\sum_{s=1}^{m} a_s z_s\right)^2 = \phi(X),$$

where $\phi(X)$ is a Cobb-Douglas production function. This technology does not exhibit constant returns to scale, since doubling the x's does not permit a doubling of the z's. With constant returns to scale, however, relative commodity prices are independent of the z's if and only if absolute commodity prices are independent of the z's. In the light of the above theorem, this assertion is equivalent to the following:

Theorem: Suppose the household's technology exhibits constant returns to scale. Relative commodity prices are independent of the commodity bundle consumed by the household if and only if the household's technology exhibits no joint production.

---

[6] Muellbauer (1973) proves this result and discusses its implications for using the household production function model as a foundation for a theory of hedonic price indexes.

[7] This was pointed out to us by John Muellbauer and Stephen A. Ross.

Proof: The "if" assertion is trivial. To establish "only if," suppose that relative prices are independent of Z. Then the locus of z's satisfying the feasibility constraint implied by (P, μ) is given by

$$z_1 = \alpha^0(P, \mu) - \sum_{s=2}^{m} \alpha^s(P, \mu) z_s.$$

By constant returns to scale, if the commodity vector Z is feasible at (P, μ), then λZ is feasible at $(P, \lambda\mu)$. Hence

$$\lambda z_1 = \alpha^0(P, \lambda\mu) - \sum_{s=2}^{m} \alpha^s(P, \lambda\mu) \lambda z_s$$

for all $\lambda > 0$. Setting $\lambda = 1/\mu$ and solving for μ yields

$$\mu = \frac{z_1}{\alpha^0(P, 1)} + \sum_{s=2}^{m} \frac{\alpha^s(P, 1)}{\alpha^0(P, 1)} z_s.$$

Hence, the cost function is of the form

$$C(P, Z) = \sum_{s=1}^{m} C^s(P) z_s,$$

and the household's technology exhibits no joint production. Q.E.D.

For the remainder of this section, we shall assume that the household's technology exhibits constant returns to scale. We have already seen that it is possible to combine the two types of constraints facing the household—the market constraint represented by the budget line and the technological constraint represented by the household production function—into the single constraint. To do this, we focus on the feasible set in the commodity space determined by goods prices, the household's technology, and its income. In the traditional theory of the household which does not recognize commodities, the boundary of the feasible set in the goods space is a hyperplane whose slope reflects the ratio of goods prices. In the household production function model with no joint production, the boundary of the feasible set in the commodity space is a hyperplane whose slope reflects both goods prices and the household's technology. As long as the boundary of the feasible set is a hyperplane, the slope corresponding to one commodity bundle on the frontier is the same as that corresponding to any other efficient bundle.

Consequently, these slopes or implicit commodity prices not only reflect the constraints facing the household but also code all of the relevant information about these constraints; no further information about technology or goods prices is needed to define the feasible set. Confronting this feasible set with preferences, the household chooses the best attainable commodity bundle. Since commodity prices are independent of the commodity bundle consumed, we can view the household as adjusting its consumption so as to bring marginal rates of substitution into alignment with the commodity price ratios determined by goods prices and technology.

With joint production (and constant or nonconstant returns to scale) the boundary of the feasible set is no longer a hyperplane. Hence, one must be very cautious in focusing attention on the slope of the boundary of the feasible set at the optimal point, since, in this case, the slope does not code all the relevant information about the constraint faced by the household. Specifically, commodity prices reflect the household's preferences as well as goods prices, income, and the household's technology. Although commodity price ratios are equated to marginal rates of substitution in equilibrium, the adjustment process takes place on both sides; as the commodity bundle changes, so do both commodity prices and marginal rates of substitution.

To illustrate the difficulties caused by the dependence of commodity prices on the commodity bundle consumed, consider two households with identical technologies who face the same goods prices and have equal incomes. These two households have the same cost function and the same feasible set in the commodity space. If the technology exhibits joint production, then the frontier of the feasible set is nonlinear and commodity prices vary with the commodity bundle chosen. Households with different tastes will select different commodity bundles, and, since the frontier is nonlinear, the commodity bundles they select will imply different commodity prices. The unwary economist might attribute some part of the difference in the consumption patterns of our two households to these differences in commodity prices, but such an interpretation would be highly misleading; the differences in commodity prices are a reflection of differences in tastes, not of differences in opportunities.

The dependence of commodity prices on the commodity bundle consumed makes it impossible to utilize the two-stage process outlined in Section 1. Although the estimation of the household's technology is complicated by joint production, it is still possible. However, the analysis of the demand for commodities in terms of maximizing a utility function subject to a linear budget constraint is no longer appropriate; commodity prices depend on the commodity bundle consumed and, hence, on the household's tastes.

If commodity prices are not analytically useful for studying the demand for commodities when the household's technologies exhibit joint production, where does this leave the household production function model? There are two alternatives. First, one can restrict the application of the model to those cases in which the technology exhibits no joint production. Although no one can object in principle to restricting the model to those cases, the usefulness of the restricted model depends on the prevalence of joint production. In Section 4 we argue that joint production is pervasive when the role of time in the household production process is recognized; this implies that restricting the application of the model in this way is very undesirable. Second, one can argue that commodity prices are not central to the household production function approach. Without commodity prices, we can derive commodity demand functions which depend on goods prices and income. These demand functions are of substantial interest even in the no-joint-production case, since we are often interested in the effects of goods prices on the consumption of commodities; but they must assume a primary role when commodity prices depend on the commodity bundle consumed.

4. The Role of Time

The allocation of time has played a central role in most applications of the household production function model. In Becker's exposition, the household purchases goods and combines them with time in a household production function to produce commodities. These commodities, rather than the goods or time, are the sources of utility.[8]

In Section 4*A* we introduce time into the basic model with constant returns to scale and no joint production. We assume that the household can sell "time" or "labor" on the market, and that it can do so at a single wage rate which is beyond its control and independent of the commodity bundle which it consumes.[9] Section 4*B* extends the model to include joint production and nonconstant returns to scale. We argue that when time is an input in the household production process, joint production is the rule and not an exception because the input of time into many household activities is itself a direct source of utility or disutility to members of the household engaged in the activity. This is a type of joint production, and, like any form of joint production, it implies that prices are not independent of the commodity bundle consumed.

A. Time in the Basic Model

The traditional theory deals with the labor-leisure choice by treating leisure as an argument of the utility function; we denote the utility function by $V(X, T_c)$, where T_c denotes "leisure," or, more precisely, time not spent at work. In the Becker terminology, T_c represents "time spent at consumption."[10] The demand functions are found by maximizing $V(X, T_c)$ subject to two interrelated constraints, one for goods,

---

[8] The notion of commodity prices involving the wage rate are anticipated in the work of Mincer (1962, 1963). Although Mincer focuses on goods and leisure (or work) rather than commodities, he writes the price of a good as the sum of a direct component related to the market price of the good and an indirect component related to the opportunity cost of time.

[9] In Becker's model (1965) there are many wage rates, and they depend on the commodity bundle the household consumes; also see Diewert (1971) for further work in this direction. We ignore these complications to focus on others.

$$\sum_{k=1}^{n} p_k x_k = A + wT_L,$$

and the other for time,

$$T_c + T_L = T,$$

where A is nonlabor income, w the fixed wage rate, T_L working time, and T total time available—for example, 24 hours, if the period is 1 day and the household consists of one person. Collapsing these two constraints into a single resource constraint by solving for T_L yields

$$\sum_{k=1}^{n} p_k x_k + wT_c = A + wT = S',$$

where S' is "full income."[11]

Maximizing $V(X, T_c)$ subject to this constraint yields demand functions for goods and a demand function for "leisure":

$$\begin{aligned} x_i &= h^i(P, w, A) \\ T_c &= h^{Tc}(P, w, A). \end{aligned}$$

These demand functions for goods and leisure are homogeneous of degree 0 in (P, w, A) and satisfy the budget constraint and the appropriate versions of the Slutsky sign and symmetry conditions. If we wish to emphasize the role of full income, we can write the demand functions as

$$\begin{aligned} x_i &= \hat{h}^i(P, w, S') \\ T_c &= \hat{h}^{Tc}(P, w, S'). \end{aligned}$$

The traditional theory is often presented in a form which ignores the labor-leisure choice and focuses on the allocation of expenditure among goods.[12] This form of the traditional theory is a special case of the full traditional model which arises when goods are separable from leisure in the household's utility function:

$$V(X, T_c) = W[\hat{V}(X), T_c].$$

In the separable case, the allocation of total expenditure is independent of the wage rate. That is, the demand functions for goods are of the form

$$x_i = h^i(P, \mu),$$

---

[10] Following Becker, we do not distinguish between "time spent at consumption" and "time spent at household production."

[11] Following Becker's usage, when there is a single wage rate independent of the commodity bundle consumed, we define full income to be $A + wT$.

[12] This issue is discussed by Mincer (1963).

where μ denotes total expenditure on goods. Total expenditure on goods is a function of all goods prices, nonlabor income, and the wage rate; but the wage rate and nonlabor income influence the demand for goods only through their influence on total expenditure. Thus, *offsetting changes in the wage rate and nonlabor income which leave total expenditure on goods unchanged will not change the goods consumption pattern.*

To incorporate time into the household production function model, we let t_r denote time devoted to the production of commodity r and $z_r = f^r(Y_r, t_r)$ denote the production function for commodity r. To find the commodity demand functions when the household's technology exhibits constant returns to scale and no joint production, we define commodity prices $\pi = (\pi_1, \ldots, \pi_m)$ to be the implicit or shadow prices implied by the technology. With constant returns to scale and no joint production, the price of each commodity is equal to its unit cost function

$$\pi_r = C^r(P, w, 1).$$

These commodity prices are independent of the commodity bundle consumed, and we can write the budget constraint in the commodity space as

$$\sum_{s=1}^{m} \pi_s z_s = A + wT = S'.$$

Maximizing the utility function $U(Z)$ subject to this constraint yields the commodity demand functions

$$z_i = g^i(\pi, w, A)$$

or, following Becker and emphasizing the role of full income,

$$z_i = \bar{g}^i(\pi, S').$$

To find the demand functions for goods implied by the household production function model, we translate the utility function $U(Z)$ into the goods space. With each vector (X, T_c) is associated a feasible set of commodity vectors which we denote by $S(X, T_c)$. We define a utility function in the (X, T_c) space, $V(X, T_c)$, by

$$V(X, T_c) = \max U(Z)$$

subject to $z \in S(X, T_c)$.[13] Maximizing $V(X, T_c)$ subject to the resource constraint yields demand functions

$$x_i = h^i(P, w, A)$$

$$T_c = h^{T_c}(P, w, A).$$

---

[13] Since the production set Ω is convex, the quasi-concavity of $U(Z)$ implies the quasi-concavity of $V(X, T_c)$.

These demand functions for goods and leisure have all the familiar properties of demand functions for goods and leisure; they are homogeneous of degree zero in (P, w, A), they satisfy the budget constraint, and they satisfy appropriate versions of the Slutsky sign and symmetry conditions.[14]

B. Complications

We saw in Section 3 that joint production and nonconstant returns to scale substantially complicate the commodity demand functions but not the goods demand functions. The same results hold when time is included in the model. The analysis of the goods demand functions earlier in this section made no use of the assumptions of constant returns to scale or no joint production and hence requires no modification when these assumptions are dropped.[15] The derivation of the commodity demand functions, however, did make use of these assumptions. Without them, the boundary of the feasible set in the commodity space is given by the cost function $C(P, w, Z) = \mu$, where

$$C(P, w, Z) = \min \sum_{k=1}^{n} p_k x_k + wT_c$$

subject to the technological constraint $Z \in S(X, T_c)$. Corresponding to each commodity bundle we define implicit or shadow commodity prices as the partial derivatives of the cost function

$$\pi_r(P, w, Z) = \frac{\partial C(P, w, Z)}{\partial z_r}$$

and implicit income by

$$I(P, w, Z) = \sum_{s=1}^{m} \pi_s(P, w, Z) z_s.$$

Unless the technology exhibits constant returns to scale and no joint production, *implicit commodity prices depend on the commodity bundle consumed*; without constant returns to scale, so does implicit income.[16]

---

[14] As Becker (1965) has emphasized, this formulation is symmetrical in its treatment of goods and time costs of commodities, placing the indirect or time costs on the same footing as the direct or goods costs. The role of the wage rate, however, appears to be more complex than those of goods prices because it enters both the cost function $C(P, w, Z)$ and the resource constraint $A + wT$, since the wage rate is both a "goods price" and a determinant of "full income." Actually, the symmetric treatment of goods and time can be completed by considering a household which has an initial endowment of goods as well as time. This yields a constraint which treats goods and time symmetrically on the income side as well as the cost side and emphasizes the formal similarity between endowments of goods and endowments of time.

[15] This is not quite true. A quasi-concave utility function in the goods-time space is compatible with only limited increasing returns to scale.

[16] In some studies, I rather than S' is defined as "full income"; see, for example, Willis (1973).

As in Section 3, we do not believe that commodity demand functions whose arguments are commodity prices which change with the commodity bundle chosen are useful analytical constructs. These "demand functions" correspond to consumer behavior under monopsony or tie-in sales, but not to the traditional theory of the household as a price taker. Consequently, the relevancy of the household production function approach to the allocation of time depends crucially on the assumptions of constant returns to scale and no joint production. Under quite general conditions, however, these assumptions are likely to be violated.

The major problem in studying the allocation of time in the household production function model centers on joint production rather than nonconstant returns to scale. Casual empiricism suggests to us that many household production processes exhibit increasing or decreasing returns to scale in ways which are directly related to the use of time. Examples related to set-up time, time spent assembling materials, etc., are easy to find but depend on the particular commodity or production process being considered. Hence, it is difficult to make a general argument concerning the importance of nonconstant returns.[17] With joint production, however, it is possible to make a general argument. The point is best approached through an example.

Consider a model with two household production processes. For definiteness, suppose the first produces a "home cooked meal" and the second a "clean house." Each of these processes requires inputs of both purchased goods and time. The time used in these activities can be that either of household members or of others hired to engage in these activities. There is no need to modify the model to include the services of hired outsiders; they are already included among the goods and their price is the relevant wage rate. To simplify the discussion, suppose the household has only one member. Formally, we can use the household production function model to represent the situation just described. There are two commodities, z_1 and z_2, which enter the household's utility function, $U(z_1, z_2)$. Each is produced by a separate production process. These processes combine market goods, including the services of hired workers, with household time to produce the commodities $z_1 + f^1(Y_1, t_1)$ and $z_2 = f^2(Y_2, t_2)$.

There is no joint production in this model, so there is no difficulty treating the demand for commodities as functions of commodity prices; however, we do not believe that the formal model is a plausible representation of the situation we have described. In particular, we object to the implied but crucial assumption that time spent cooking and time spent cleaning are "neutral" from the standpoint of the household and that only the "outputs" of these production processes enter the household's utility function. A more plausible assumption is that the household is not indifferent among all situations which involve the same output of home cooked meals and clean houses but involve different amounts of hired labor and household labor. Instead, we suggest that household time spent cooking or cleaning is a direct source of utility or disutility to the household. Consequently, household decisions about the allocation of time reflect not only production considerations but also direct household preferences as to the uses of time.

---

[17] Muth (1966) argues that constant returns to scale are a tautological property of any production process. Even if one accepts this as a metaphysical proposition, it has no relevancy here; what we require is constant returns to scale in (Y_r, t_r), the inputs which the household can purchase on the market, not constant returns to scale in some augmented set of inputs.

Formally, the situation can be treated in the household production function framework, but since household time spent cooking and cleaning is a source of utility, there are now four commodities instead of two. The household's utility function is then $U(z_1, z_2, z_3, z_4)$. We still have two production processes or activities which combine goods (including the services of nonmembers of the household) and household time to produce commodities, but these processes now involve joint production;

$$z_1 = F^1(Y_1, t_1)$$
$$z_3 = t_1$$

and

$$z_2 = F^2(Y_2, t_2)$$
$$z_4 = t_2.$$

The commodities z_3 and z_4 reflect the household time spent in the production activities associated with z_1 and z_2, respectively.

With joint production of the type we have specified, the cost function $C(P, w, Z)$ assumes the form

$$C(P, w, Z) = C^1(P, w, z_1, z_3) + C^2(P, w, z_2, z_4),$$

where the superscripts on the cost functions identify the process or activity rather than the commodity. Since the implied commodity prices depend on the commodity bundle consumed, they reflect the household's tastes as well as goods prices and the household's technology.

In production theory an entrepreneur has no inclination to devote homogeneous labor to one activity rather than another except in order to maximize the value of his output. In the household, time spent in various activities is often a direct source of satisfaction or dissatisfaction as well as an input into a production activity. This is a case of joint production, and the household production function approach is not a suitable framework for analyzing joint production. We conclude that the household production function model provides a satisfactory account of the allocation of time only for households which are indifferent among alternative allocations of their time.

5. Alternative Approaches

When commodity prices depend on household tastes, we need an alternative to studying the demand for commodities as functions of commodity prices. One possibility is to regard the demand for commodities as functions of goods prices, the wage rate, nonlabor income, and the household's technology. This provides a suitable framework for analyzing the demand for commodities when the household's technology exhibits joint production or nonconstant returns to scale. It retains the household production function model's emphasis on commodities while dispensing with commodity prices. But this approach is appropriate only when studying variables which are the outputs of production processes. When household activities can more plausibly be interpreted as directly producing a specific type of utility or satisfaction, the appropriate alternative is to analyze the "inputs"—the allocation of goods and time among household activities—as a function of goods prices, the wage rate, and nonlabor income.

A. Commodities as Functions of Goods Prices

Even with joint production or nonconstant returns, it is possible to estimate the household's technology and construct commodity prices. The commodity prices, however, depend on the commodity bundle consumed as well as on the goods prices and the household's technology. In this case, we believe that it is best to dispense with commodity prices as arguments of the demand function for commodities and to treat the demand for commodities as a function of goods prices, wage rates, and nonlabor income. This has the effect of making the household's technology a parameter of the demand functions, but we believe this is preferable to building tastes into commodity prices. First, it is a conceptual advantage to define prices so that they reflect the constraints facing the household, but not its tastes. This is consistent with the traditional role of prices in demand analysis, where they represent constraints on feasible behavior and are independent of tastes. Second, prices which depend on tastes are not particularly useful analytical constructs. The theorems of traditional demand theory presuppose that the household is a price taker.

Formally, maximizing $U(Z)$ subject to the constraint $C(P, w, Z) = A + wT$ yields demand functions which we denote by $Z = g(P, w, A)$. These demand functions show the household's consumption of commodities as functions of goods prices, the wage rate, and nonlabor income instead of as functions of commodity prices and implicit income. Technical progress—changes in the household's technology—as well as taste change will be reflected by shifts in these demand functions. Because there is no natural correspondence between goods prices and commodity quantities, the Slutsky sign and symmetry conditions have no meaning for the demand functions $Z = g(P, w, A)$.

The $Z = g(P, w, A)$ demand functions reflect the influence of both technology and tastes, and, without explicit estimates of the technology, we cannot tell whether variations in consumption are due to variations in tastes or to variations in technology. To argue that all variations in consumption behavior should be attributed to variations in technology because we have a better theory of technology than of tastes is logically unconvincing. Furthermore, although it may be possible to explain consumption behavior in terms of either taste change or technical progress, the welfare implications are quite different.

B. Commodities—Technology versus Tastes

In the previous section, we suggest that in the presence of joint production or nonconstant returns, the demand for commodities be treated as a function of goods prices rather than commodities prices. In this section, we discuss the case of variables which should not be interpreted as commodities and which are not suitable as dependent variables of demand functions.

A fundamental problem in many applications of the household production function model is the need to assume that household activities being investigated are production processes. If a household activity is a production process, we call the outputs commodities. For example, the commodity "heat" is produced by a furnace and fuel. The relevance of technology for some situations, however, does not imply its relevance for all. Specifically, in some potential applications, the variables being investigated are not "commodities" (the outputs of production processes) but rather utilities (numbers representing preference orderings). These misapplications confound tastes with technology by interpreting specific utility functions as production functions.[18]

Our point is essentially an applied rather than a theoretical one. Consider, for example, the notion of "child quality." One approach to child quality in terms of the household production function model is to assume that market goods and time produce (perhaps jointly) a number of *narrowly* defined child-related commodities. The vector of these narrowly defined commodities may include variables such as verbal, mathematical, and artistic skills, grades in school, and even the negative of the number of cavities. A second approach, again using the household production function model, is to treat "child quality" as itself a commodity produced by the household's technology. We believe that the aggregation of the narrowly defined child-related commodities into a single index of child quality reflects the preference ordering of the household and not its technology.[19]

---

[18] It might be thought that this difficulty could be described in terms of "measurability"; that is, commodities are cardinally measurable while utilities are ordinally measurable. This, however, is not the case. Even the discovery of a unique cardinal measure of utility would not alter our objection to the confounding of tastes and technology in the household production function model. We believe, however, that some thought should be given to the status of ordinal variables as inputs or outputs in production theory, especially when unique indices are not widely accepted.

In the household production function model as well as in the traditional demand theory approach, the household is maximizing utility rather than output. In applying production theory to the household, the household production function literature has not attempted to draw the line indicating where production processes stop and utility begins. In empirical work, where direct measures of the relevant variables are not available, the two are easily confused.

For example, suppose that the household's utility function is separable, the goods are partitioned into subsets such as food, clothing, and housing, and the utility function is of the form

$$V(X) = W[V^1(x_1), \ldots, V^m(x_m)],$$

where the V's are "specific utility functions." This is similar to a household production function model in which no good is used in the production of more than one commodity; it differs from the household production function model only because the "outputs" of the specific utility functions are not measurable.[20] Hence, in the absence of direct measures of the commodities produced by the household production function, it may be difficult to distinguish between a production function and a specific utility function interpretation of the household's behavior.

Even if we have no observations on the production of commodities by the household, we can still study the allocation of time and goods among household activities. Although Becker, in his seminal article "A Theory of the Allocation of Time," concentrates on household production functions and the resulting commodities, a study of the underlying activity or input demand functions provides a feasible alternative.[21]

---

[19] Recent papers which discuss the quality of children as an output of a production process include DeTray (1973), Michael (1973), and Willis (1973). For a criticism of this approach similar to ours, see Reid (1973). Lindert (1973) presents an alternative approach which foregoes calculating the *price* of a child and instead concentrates on actual *cost* data from budget studies. He argues that the price of a child includes psychic value and is thus not useful for empirical work. See also Gronau (1973).

[20] Muth (1966, p. 700) suggests that the household production function model yields "a utility function which is weakly separable when viewed as a function of commodities [e.g., goods] purchased on the market." Nerlove (1974) quotes this with approval, while Michael and Becker (1973) are more cautious, suggesting that the model implies that the utility function "is separable in *the goods and time used to produce a given commodity*" (italics are ours). We agree with Muellbauer (1973), who points out that separability is an extra assumption, not a consequence of the basic household production function model. A good like electricity may be used in the production of many commodities without implying joint production or nonconstant returns. The assertion made by Michael and Becker is technically correct, but the implication of their observation is that electricity used for cooking and electricity used for heating must be treated as two distinct goods.

To study the allocation of goods and time within the household, we assume that the household allocates both goods and time among activities. Suppose that there are l activities ,and we let Y_r denote the vector of goods devoted to activity r and t_r denote the household time devoted to activity r. We let (Y, t) denote the vector specifying the allocation of goods and household time among activities. In the household production function model there is a vector of commodities associated with each activity; in the example involving cooking and cleaning each activity produced two commodities. The utility function $U(Z)$ enables us to assign a utility to each activity-input vector,

$$W(Y, t) = U(Z),$$

where Z is the commodity vector produced by (Y, t). Maximizing this *utility function subject to the budget constraint*

$$\sum_{s=1}^{l}\sum_{k=1}^{n} p_k y_{sk} + w\sum_{s=1}^{l} t_s \leq A + wT$$

implies a system of demand functions for activity inputs. We denote these input demand functions by

$$\begin{aligned} y_{ri} &= F^{ri}(P, w, A) \\ t_r &= F^{tr}(P, w, A) \end{aligned}$$

or, in vector form, $(Y, t) = F(P, w, A)$.

Even if there are no observable outputs, we can still derive the demand functions for the activity inputs as functions of P, w, and A from the utility function $W(Y, t)$.[22] The existence of demand functions for the y_{ri} variables, however, does depend on our ability to distinguish the use of good i in activity r from its use in other activities. Similarly, in order to study the allocation of time, it must be possible to divide the day into nonoverlapping activities. If this can be done, then the allocation of time and goods among activities can be analyzed without the household production function apparatus. In any case, the above condition is also necessary for the household production function approach to be relevant. Hence, a study of the allocation of time and goods among activities requires fewer assumptions than does the study of commodities. In particular, it avoids the need to assume that a commodity is actually being produced by a technological process.

---

[21] One can study the allocation of time without making use of household production function constructs (see Mincer 1962, 1963). Although his theoretical discussion combines goods prices and wages into a commodity price construct, his empirical work enters these prices separately.

[22] Becker's analysis does not distinguish sharply between the household's consumption of commodities and the allocation of goods and time within the household. In part this reflects his assumption that commodities are produced by a fixed-coefficient technology, so that household time devoted to the production of commodity r is proportional to the household's output of commodity r. When the technology exhibits fixed coefficients, it is possible to use the input of household time as a proxy for the output of the commodity; but when the technology permits substitution of goods or hired labor for the time of household members, then it is necessary to distinguish between the input of household time and the output of commodities.

6. Conclusions

In this paper we have provided a critique of the new theory of household behavior based on household production functions, and, in particular, of its application to problems involving the allocation of time. Where the new theory does not provide a viable model, we have suggested alternative approaches.

The application of the household production function approach requires constant returns to scale and the absence of joint production. If these conditions are not satisfied, commodity prices depend on the household's preferences, and therefore fail to serve the traditional role of prices in consumer theory; that is, they do not code all the relevant information about the constraints faced by the household. Hence, if the household's technology exhibits nonconstant returns to scale or joint production, demand functions based on commodity prices are seriously misleading.

When the production of a commodity involves inputs of the household's time, the production process is likely to exhibit joint production. This is because the household derives utility or disutility from the time it devotes to each activity as well as from the nominal "commodity output" of the activity. Technically, the household time devoted to an activity is both an input and an output, and, if the activity yields other commodity outputs as well, the activity exhibits joint production. When the technology exhibits joint production or nonconstant returns to scale, we suggest that the demand for commodities be analyzed in terms of goods prices rather than commodity prices.

The household production function approach has been applied to cases in which the variable under study may be more appropriately viewed as a utility than as the output of a production process. If commodities are not being produced, it makes no sense to analyze the demand for them; but we can, however, examine the household's allocation of goods and time among activities as a function of goods prices and the wage rate, provided only that these allocations are themselves observable.

References

Becker, Gary S. "An Economic Analysis of Fertility." In *Demographic Change in Developed Countries.* Universities—National Bureau Conference Series 11. Princeton, N.J.: Princeton Univ. Press, 1960.

———. "A Theory of the Allocation of Time." *Econ. J.* 75, no. 299 (September 1965): 493–517.

DeTray, Dennis N. "Child Quality and the Demand for Children." *J.P.E.* 81, no. 2, pt. 2 (March 1973): 70–90.

Diewert, W. E. "Choice on Labour Markets and the Theory of the Allocation of Time." Department of Manpower and Immigration, Ottawa, 1971.

———. "Applications of Duality Theory." Discussion Paper no. 89, Univ. British Columbia, 1972.

Gronau, Reuben. "The Effect of Children on the Housewife's Value of Time." *J.P.E.* 81, no. 2, pt. 2 (March 1973): 168–99.

Hall, Robert E. "The Specification of Technology with Several Kinds of Output." *J.P.E.* 81, no. 4 (July 1973): 878–92.

Lancaster, Kelvin J. "A New Approach to Consumer Theory." *J.P.E.* 74, no. 2 (April 1966): 132–57. (*a*)

———. "Change and Innovation in the Technology of Consumption." *A.E.R.* 56, no. 2 (May 1966): 14–23. (*b*)

———. *Consumer Demand: A New Approach.* New York: Columbia Univ. Press, 1971.

Lindert, Peter H. "The Relative Cost of American Children." Economic History Discussion Paper, Eh73-18, Univ. Wisconsin, March 1973.

Michael, Robert T. "Education and the Derived Demand for Children." *J.P.E.* 81, no. 2, pt. 2 (March 1973): 128–64.

Michael, Robert T., and Becker, Gary S. "On the New Theory of Consumer Behavior." *Swedish J. Econ.* 75 (1973): 378–96.

Mincer, Jacob. "Labor Force Participation of Married Women." In *Aspects of Labor Economics*, edited by H. Gregg Lewis. Universities—National Bureau Conference Series 14. Princeton, N.J.: Princeton Univ. Press, 1962.

———. "Market Prices, Opportunity Costs, and Income Effects." In *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, edited by C. Christ et al. Stanford, Calif.: Stanford Univ. Press, 1963.

Muellbauer, John. "Household Production Theory, Quality and the 'Hedonic Technique.'" *A.E.R.* (in press).

Muth, Richard F. "Household Production and Consumer Demand Functions." *Econometrica* 34, no. 3 (July 1966): 699–708.

Nerlove, Marc. "Household and Economy: Toward a New Theory of Population and Economic Growth." *J.P.E.* 82, no. 2, pt. 2 (March/April 1974): 200–218.

Reid, Margaret G. "Comment." *J.P.E.* 81, no. 2, pt. 2 (March/April 1973): 165–67.

Willis, Robert J. "A New Approach to the Economic Theory of Fertility Behavior." *J.P.E.* 81, no. 2, pt. 2 (March/April 1973): 14–65.

Capital-Skill Complementarity, Income Distribution, and Output Accounting

---

P. R. Fallon and P. R. G. Layard

Centre for the Economics of Education, London School of Economics and Political Science

This article presents evidence for the view that physical capital is more complementary to educated labor than to less educated labor. For this reason previous estimates of elasticities of substitution between educated and less educated labor are too high. Using a two-level CES production function, we run international cross-sectional regressions at the level both of the economy and of individual sectors. The whole-economy model allows for the effects of wages on educational choices as well as vice versa. It predicts that, as total capital per head rises, the share of physical capital in national income falls and that of human capital rises. The production function also shows that intercountry differences in output per head are due more to differences in physical capital than in human capital.

Introduction and Summary

How quickly do the returns to education fall when the number of educated people rises? This has been a crucial question for the philosophy of educational planning, since the case for manpower forecasting and planning is stronger the less easy the process of substituting educated for less educated people.

In answer to the question, Blaug (1967) and others pointed out that U.S. rates of return to education had been remarkably constant over time despite a vast increase in the educated labor force. This suggested that substitution was relatively easy. Likewise, cross-sectional data on countries (Bowles 1970) and on U.S. states (Dougherty 1972) showed that the relative wages of the educated tend to vary with their relative numbers

We are grateful for financial support to the U.K. Social Science Research Council, and to C. R. S. Dougherty and G. Pyatt for helpful comments. For further discussion of many of the topics see Fallon (1974).

[*Journal of Political Economy*, 1975, vol. 83, no. 2]

© 1975 by The University of Chicago. All rights reserved.

by only small amounts—evidence again, it was claimed, of easy substitution.

However, the inference, whether from time series or cross section, may not be valid if other things are varying at the same time. And the relative use of physical capital always varies, the ratio of physical capital to raw labor generally rising with the ratio of educated to raw labor. If physical capital is more complementary to educated than to raw labor, this could then explain why the relative wages of the educated are not much lower when their relative numbers are much higher.[1] Equally, if this is the explanation, planners should avoid excessive educational expansion not accompanied by physical investment, since this could produce a rapid fall in relative wages.

This "capital-skill complementarity hvpothesis" has been advanced by Griliches (1969, 1970) and partially confirmed on a cross section of U.S. states. Because of lack of data, he assumed constancy across states in the absolute wage of educated manpower and in the rate of return on physical capital in any given industry. The present study relies instead on international data on 23 countries, including detailed information on national education-specific wages and data on physical capital stocks and rentals at the national, though not at the sectoral, level. The data have the advantages of wide variation but obvious problems of comparability. As regards the form of the relationships involved, Griliches restricted these by assuming constant own- and cross-elasticities of demand in each factor demand equation, while we restrict them by assuming an explicit form of production function—the two-level CES function.[2] This has the advantage, apart from yielding meaningful results, that we can use the explicit function to account for income differences between countries and for differences in the functional distribution of income.

The paper is constructed as follows. Section I examines capital-skill complementarity and the ease of substitution between types of labor at the level of the whole economy. The problem here is that factor prices affect educational choices as well as vice versa, and a complete model is therefore specified. Section II examines the same question at the sectoral level, factor prices being taken here as exogenous. Both Sections I and II confirm the capital-skill complementarity hypothesis and suggest much lower elasticities of substitution between educated and raw labor than either Bowles or Dougherty. Sections III and IV revert to the whole-economy model. Section III illustrates how the production function, together with the relation determining the supply of educated people, determines the evolving pattern of income distribution as countries get richer (as measured by their total capital per head). If the supply of educated people is always such that the rates of return to education and to human capital are in fixed proportion, the parameters of the production function correctly predict that as economic progress occurs, human capital per head will grow faster than physical capital per head and the share of physical capital in national income will fall. In Section IV we examine the relative importance of physical and human capital in explaining income differences between countries and conclude (contrary to the claims of Krueger [1968]) that physical capital is generally the more important.

---

[1] Other possible hypotheses explaining the constancy of relative wages over time include the following: (*a*) Luxuries are more education-intensive than necessities. Here Folger and Nam (1967) show that shifts in the sectoral composition of the labor force account for a negligible proportion of the overall rise in relative proportions of educated people in the United States. (*b*) Technical progress is biased (in the Hicksian sense) toward educated rather than raw labor. (*c*) Technical progress is fastest in education-intensive industries having demand elasticities greater than unity. Both (*b*) and (*c*) are extremely difficult to study. Other possible hypotheses explaining the cross-sectional stability of relative wages are: (*a*) as above; here Fallon (1974, chap. 8) shows the same as Folger and Nam's time series; (*b*) the factor-price equalization theorem; however, this implies identical factor proportions in all countries in any one industry, which is inconsistent with the data used in our Section II.

[2] We also tried the Griliches approach (modified so that prices are the dependent variables and quantities the independent variables), but found that the results for the richer and poorer countries taken separately were grossly inconsistent with each other and with the results for the pooled sample, whether judged by *F*-tests or by consistency of the signs of the coefficients.

I. The Pattern of Substitution: Whole-Economy Level

The Problem

Suppose there are only three factors of production, physical capital (K), skill (S—to be defined), and other labor (N—to be defined), and output (Y) is determined by: $Y = f(K, S, N)$. Then the capital-skill complementarity hypothesis, in the form in which we are interested, states that:

$$\frac{\partial}{\partial_K}\left(\frac{f_S}{f_N}\right) > 0,$$

which implies

$$\frac{f_{SK}}{f_S} > \frac{f_{NK}}{f_N}.$$

By contrast, the production function used by Bowles and Dougherty assumes that physical capital has no effect on the relative marginal products of skill and other labor, the function having the form

$$Y = f(K, [bS^{\theta} + (1 - b)N^{\theta}]).$$

To examine the capital-skill complementarity hypothesis we need to use a form of function which permits it to be confirmed or rejected. The

Cobb-Douglas does not, and the simplest form which does is the two-level CES function,

$$Y = A[aQ^{\rho} + (1 - a)X_1^{\rho}]^{1/\rho} \qquad (\rho \leq 1),$$

where

$$Q = [bX_2^{\theta} + (1 - b)X_3^{\theta}]^{1/\theta} \qquad (\theta \leq 1),$$

and X_1, X_2, and X_3 are any permutation of K, S, and N.[3] The Bowles function would correspond to the version where K played the role of X_1 ("the odd man out") and was equally complementary to skill and other labor. Only if S or N played the role of X_1 could capital be more (or less) complementary to skill than to other labor. We have tested all three permutations of the two-level function using the two-stage least-square models developed later and found that the production function estimates where N played the role of X_1 were the most satisfactory on a number of criteria.[4]

Our preferred functional form is thus

$$Y = A[aQ^{\rho} + (1 - a)N^{\rho}]^{1/\rho} \qquad (\rho \leq 1),$$

where $\qquad\qquad (1)$

$$Q = [bK^{\theta} + (1 - b)S^{\theta}]^{1/\theta} \qquad (\theta \leq 1).$$

If the capital-complementarity hypothesis is true, $\rho > \theta$, and vice versa.[5] In other words, the direct elasticity of substitution between N and Q (i.e., $1/[1 - \rho]$) must exceed the direct elasticity of substitution, within the "nest," between K and S (i.e., $1/[1 - \theta]$).

This is really all that needs to be said about the capital-skill complementarity hypothesis given the two-level CES function However, it may be useful to relate the condition $\rho > \theta$ to the other elasticities used in general discussions of the issue. Of these, the most relevant is the elasticity of complementarity (Hicks 1970). This is defined for a constant-returns production function $f(X_1, \ldots, X_n)$ as

$$c_{ij} = \frac{1}{v_j}\frac{\partial \log f_i}{\partial \log X_j}\left(= \frac{1}{v_i}\frac{\partial \log f_j}{\partial \log X_i} = c_{ji}\right),$$

where v_j is the share of the jth factor in output, and $\partial \log f_i/\partial \log X_j$ indicates the proportional effect on the marginal product of the ith factor of a change in the quantity of the jth factor, holding all other input quantities constant. Clearly

$$c_{SK} = \frac{1}{v_K}\cdot\frac{f_{SK}}{f_S}\cdot K$$

---

[3] Another function which does is the log-quadratic, but this does not enable one to take advantage of data on relative prices.

[4] The evidence is presented in Appendix A.

[5] Taking derivatives in (1), if $(f_{SK}/f_S) > (f_{NK}/f_N)$,

$$a(1 - a)b(1 - b)A^{2\rho}Y^{2(1-\rho)}K^{\theta-1}S^{\theta-1}N^{\rho-1}Q^{\rho-2\theta}(\rho - \theta) > 0.$$

and

$$c_{NK} = \frac{1}{v_K} \cdot \frac{f_{NK}}{f_N} \cdot K,$$

so that our form of the capital-skill complementarity hypothesis is fulfilled if $c_{SK} > c_{NK}$. In the case of our two-level CES function,[6]

$$c_{SK} = 1 - \rho + \frac{1}{v_Q}(\rho - \theta)$$

and

$$c_{NK} = c_{SN} = 1 - \rho.$$

This confirms that the capital-skill complementarity hypothesis requires that $\rho > \theta$. We may also note that, in Hicks's language, the pairs N, K and N, S are q-complements whatever the value of ρ (< 1), while capital and skill (K and S) may or may not be q-complements but must be if $\rho > \theta$.

From our point of view, the elasticity of complementarity is a more useful concept than the better-known Allen elasticity of substitution, used in the Griliches (1969) analysis. The reason is that the former deals with the effect of a factor quantity change on factor prices (other factor quantities and output price constant), while the latter deals with the effect of a factor price change on factor quantities (other factor prices and output quantity constant). The Allen elasticity of substitution is defined as

$$\sigma_{ij} = \frac{1}{v_j} \cdot \frac{\partial \log X_i}{\partial \log p_j} \left(= \frac{1}{v_i} \cdot \frac{\partial \log X_j}{\partial \log p_i} = \sigma_{ji} \right).$$

However, under constant returns to scale it can be shown that, even in the n-factor case,[7] $c_{SK} > c_{NK}$ implies and is implied by $\sigma_{SK} < \sigma_{NK}$. In our two-level CES function

$$\sigma_{SK} = \frac{1}{1 - \rho} + \frac{1}{v_Q}\left(\frac{1}{1 - \theta} - \frac{1}{1 - \rho}\right)$$

and

$$\sigma_{NK} = \sigma_{SN} = \frac{1}{1 - \rho}.$$

This confirms again that the capital-skill complementarity hypothesis requires $\rho > \theta$. In addition we may note that, in Hicks's language, the pairs N, K and N, S must be p-substitutes, while capital and skill (K and S) may or may not be p-substitutes but must be if $\rho < \theta$.

---

[6] $c_{SS} = 1 - \rho + \frac{1}{v_Q}(\rho - \theta) + \frac{1}{v_S}(\theta - 1) < 0$ (and analogously for c_{KK});
$c_{NN} = 1 - \rho + \frac{1}{v_N}(\rho - 1) < 0.$

[7] On the relationship between elasticities of complementarity and substitution, see Sato and Koizumi (1973).

So much for capital-skill complementarity—that is, the effect of changes in *capital* on rates of return to education. Turning to the effect of *education* on rates of return to education, the measure normally used is the direct elasticity of substitution, defined as

$$d_{ij} = -\frac{\partial \log (X_i/X_j)}{\partial \log (f_i/f_j)} \left(= -\frac{\partial \log (X_j/X_i)}{\partial \log (f_j/f_i)} = d_{ji}\right),$$

that is, the proportional change in the relative quantities of the ith and jth factors for a given change in their relative marginal products, all other factors and output being held constant. To measure the rate at which the returns to education will fall as the relative number of educated people rises, we take the inverse

$$\frac{1}{d_{SN}} = -\frac{\partial \log (f_S/f_N)}{\partial \log (S/N)}.$$

In the case of our function,[8]

$$\frac{1}{d_{SN}} = \frac{v_S v_N}{v_S + v_N} (2c_{SN} - c_{SS} - c_{NN}).$$

Like all direct elasticities, this is positive and lies between $1/(1 - \theta)$ and $1/(1 - \rho)$. It may not be strictly what is needed from a planning point of view, since, if the number of skilled people is increased in a country, the number of unskilled is constrained to decline by the same number. So output rises and substitution does not occur along an isoquant. Thus, unless the production function is homothetic in skilled and unskilled labor, the proportional change in f_S/f_N for a change in S/N such that $dS + dN = 0$ slightly exceeds $1/d_{SN}$ (provided $f_S > f_N$). However, in our estimated function the difference is small for most countries,[9] and we shall simply quote the direct elasticities of substitution as the occasion arises.

The Model with More Educated and Less Educated Labor

In order to estimate the parameters of the production function (1), we must specify the model generating our observations. It is convenient to begin with some definitions:

---

[8] The expression for $1/d_{KN}$ is analogous, while $1/d_{SK} = 1 - \theta$.

[9] If S' is defined as human capital (see below) and N' as the total labor force, there is no constraint of the form $dS' + dN' = 0$, and the rate at which the returns to education fall is given by

$$-\frac{d \log (f_{S'}/f_{N'})}{d \log S'},$$

which exceeds $1/d_{S'N'}$.

Quantity	Price (1963 $)	Definition of Quantity
K	r	Physical capital ($000, 1963 prices)
S	z	Workers with 8 or more years of education (000s)
N	w	Workers with 7 or less years of education (000s)
Q	q	$(bK^\theta + (1 - b)S^\theta)^{1/\theta}$
Y	...	Output ($000, 1963 prices)
P	...	Population (000s)

The following two basic estimating equations follow from the condition that price equals marginal product.

$$\log \frac{r}{z} = \log \frac{b}{1 - b} + (\theta - 1) \log \frac{K}{S}. \tag{2}$$

$$\log \frac{q}{w} = \log \frac{a}{1 - a} + (\rho - 1) \log \frac{Q}{N}. \tag{3}$$

Clearly (2) has to be estimated before (3), since it provides the parameters needed to construct the variables Q and q.

However, ordinary least-square estimates of these equations could be seriously biased if relatively more people get educated in countries where the private rate of return to education is high relative to the rate of return on physical capital. We therefore need an equation for the supply of educated labor, relating this to the relative rate of return and (owing to the consumption aspects of education) to income per head. The private rate of return to education is approximately proportional to the ratio of more educated to less educated wages (z/w), so we have

$$\log \frac{S}{N} = \log c_0 + c_1 \log \frac{z/w}{r} + c_2 \log \frac{Y}{P}. \tag{4}$$

Clearly S, N and P are not unrelated, since $S + N$ is the total labor force. There is, therefore a total labor supply equation,[10]

$$\log (S + N) = \log d_0 + d_1 \log P; \tag{5}$$

and the model is completed by the product exhaustion condition

$$Y = rK + zS + wN. \tag{6}$$

Treating Q and q not as separate variables but as weighted aggregates of K, S, r, and w, we have six endogenous variables (Y, S, N, r, z, w) and two exogenous variables (K and P). If the system were linear, we should conclude that equation (2), which includes six linear restrictions, was overidentified. The same holds for (3), as can be seen by explicitly adding two equations defining log Q and log q. Equation (4) would be just identified. However, the system is not linear; but Fisher (1966, p. 149) has shown that, in general, nonlinear systems are not less identified than the corresponding linear systems. We therefore estimate equations (2), (3), and (4) using two-stage least squares, log K and log P being the exogenous variables.

---

[10] In many ways it would have been simpler to treat the total labor supply as exogenous, but we should then have had to explain S/N in eq. (4) by income per worker rather than by income per head; the latter seems more plausible. Fortunately, the estimated value of d_1 for the whole sample was 1.03, making the solution values of real prices and quantity ratios approximately independent of the size of country and dependent only on the ratio of the two exogenous variables K and P.

The Model with Human Capital and Raw Labor

The preceding model is rather crude in the way it adds together all workers with higher education (L_H) and with secondary education (L_S) to form a single category for more educated workers (S), and all workers with primary education (L_P) and with no education (L_0) to form a single category of less educated workers (N).[11] If we want to take into account differences between countries in L_H/L_S and L_P/L_0 while continuing to work within a three-factor model, the natural approach is to differentiate between human capital (S') and raw labor (N'). The latter (N') consists simply of the number of workers in the labor force, irrespective of how much human capital, if any, each worker has; and its price (w') is the wage of workers with no education ($= w_0$). Human capital (S') is a weighted sum of the human capital formed by primary education (L_P), by secondary education (L_S), and by higher education (L_H). Since we are treating each type of human capital as a perfect substitute for each other, we need a constant set of weights, the natural weights being the average world price for the services of each type of capital. So

$$S' = (\bar{w}_H - \bar{w}_0)L_H + (\bar{w}_S - \bar{w}_0)L_S + (\bar{w}_P - \bar{w}_0)L_P,$$

where $\bar{w}_H$ is the world average wage of higher-educated workers (the unweighted country average), and likewise for the other categories. The dimension of S' is units of skill. Its price (z') equals its income divided by its quantity and is a money flow per unit of time. Thus

$$z' = \frac{(w_H - w_0)L_H + (w_S - w_0)L_S + (w_P - w_0)L_P}{S'},$$

where w_H are the wages of higher-educated people in the country in question and likewise for the other categories. This price (z') is not the same as the private rate of return to education, which depends on the return to education (z') relative to its cost, which in turn is roughly proportional to the earnings of the uneducated (w'). Thus, if all educated workers had primary education only, the rate of return would be proportional to $(w_p - w_0)/w'$, which would be proportional to z'/w'. It is this latter expression which we take as our proxy for the rate of return to education.[12]

---

[11] L_H = 13+ years of education, L_S = 8–12, L_P = 1–7, L_0 = less than 1.

We can now set up our model. The key feature is that now total capital (C), both human and physical, is taken as exogenous rather than physical capital (K) only, as in our first model. Physical capital is measured in the same way as before under the assumption that the price of a capital good is equalized across the sample by international trade. On the other hand, human capital must be measured in units of domestic value rather than, as S', in physical units. Since the monetary rate of return on education is influenced by its nonmonetary returns, it seems reasonable to take the rate of return on physical capital (r) as the rate of time preference, so that the value of human capital is $z'S'/r$. The total capital identity is then

$$C = K + \frac{z'S'}{r}. \tag{7}$$

The balance between human and nonhuman capital is determined as before by relative rates of return and income per head:

$$\log \frac{z'S'/r}{K} = \log c_0 + c_1 \log \frac{z'/w'}{r} + c_2 \log \frac{Y}{P}. \tag{4'}$$

The total labor force (N) is related to the total population (P) by

$$\log N' = \log d_0 + d_1 \log P, \tag{5'}$$

and the model is completed by equations (1′), (2′), (3′), and (6′), which are identical with (1), (2), (3), and (6) except that N', w', S', z', Q', and q' replace the corresponding variables. This model has one equation more than the first and one additional variable (C). The identification

---

[12] In the case of more than one type of education, if the relative prices of the different types of human capital are constant, as assumed in constructing S', then the relative rates of return on the different types of human capital vary. Thus suppose there are two levels of education, secondary and primary, each lasting 1 year. By assumption,

$$\frac{w_S - w_0}{w_P - w_0} = k = \frac{\bar{w}_S - \bar{w}_0}{\bar{w}_P - \bar{w}_0}.$$

By definition

$$r_P = \frac{w_P - w_0}{w_0} = \frac{w_P}{w_0} - 1$$

$$r_S = \frac{w_S - w_0}{w_P} = \frac{k(w_P - w_0)}{w_P} = k - \frac{kw_0}{w_P} = k - \frac{k}{r_P + 1}.$$

Thus r_S and r_P vary positively, r_P/r_S rising as r_P rises (an observation confirmed by Psacharopoulos 1973). Therefore

$$\frac{z'}{w'} = \frac{(w_S - w_0)L_S + (w_P - w_0)L_P}{(\bar{w}_S - \bar{w}_0)L_S + (\bar{w}_P - \bar{w}_0)L_P} \cdot \frac{1}{w_0} = \left(\frac{w_P - w_0}{\bar{w}_P - \bar{w}_0}\right) \frac{1}{w_0} = \frac{r_P}{\bar{w}_P - \bar{w}_0}.$$

So z'/w' is proportional to r_P and varies positively with r_S.

TABLE 1

SUBSTITUTION PARAMETERS: WHOLE ECONOMY

	Eqq. (2), (2′) $\theta - 1$	Eqq. (3), (3′) $\rho - 1$	$\rho - \theta$	No. of Observations
All countries:				
Using S, N	−3.45 (−1.8)	−0.67 (−11.0)	2.78 (1.5)	22
Using S', N'	−1.72 (−4.3)	−0.96 (−8.5)	0.76 (1.8)	22
Richer countries:				
Using S, N	−1.81 (−2.1)	−0.54 (−3.8)	1.27 (1.4)	9
Using S', N'	−0.94 (−4.6)	−1.27 (−5.7)	−0.33 (−1.1)	9
Poorer countries:				
Using S, N	−1.50 (−1.2)	−0.47 (−4.1)	1.03 (0.8)	13
Using S', N'	−2.03 (−1.6)	−0.41 (−2.0)	1.62 (1.3)	13

NOTE.—Figures in parentheses are t-statistics (see text re third column).

of equations (2′), (3′), and (4′) is thus the same as for (2), (3), and (4). Estimation is by two-stage least squares, log C and log P being the exogenous variables.

Estimates

The data used relate to 1963 and are described in Appendix B, while the countries included are shown in table 4. Broadly the wage data are based on the sources used in the rate-of-return studies summarized in Psacharopoulos (1973),[13] while the employment data are based on the 1961 Censuses. The capital measure is the sum of gross fixed capital formation from 1949 to 1963 inclusive, at 1963 prices, and capital rental is the ratio of estimated capital income to capital so measured.

Table 1 shows the estimates of θ and ρ provided by equations (2) and (3) (and [2′] and [3′]). The regressions are done for the whole sample and for the richer and poorer countries separately. We thus have six estimates of $\rho - \theta$, all but one positive, as predicted by the capital-skill complementarity hypothesis. To measure the significance of the differences between $\hat{\rho}$ and $\hat{\theta}$, table 1 assumes that they are independent. The t-statistics shown in the third column are computed on this assumption, but a sensitivity analysis has suggested that a weak positive relation exists between $\hat{\rho}$ and $\hat{\theta}$, in which case these t-statistics are (absolutely) too low. At the crude aggregate level there is thus some mild confirmation of the capital-skill complementarity hypothesis.

As for the substitutability of more and less educated people, for the whole sample the direct elasticity of substitution is 0.61, much less than the values of 6–8 found by Bowles. This is partly due to differences in data and partly due to differences in specification. If on our data we use the two-level function with K taking the role of X_1, which is analogous to Bowles's procedure, we find an elasticity of 3.54.[14]

---

[13] We are very grateful to the author for making the data available to us.

II. The Pattern of Substitution: Sectoral Level

Estimating Approach

Economy-wide functions are, however, necessarily crude and we turn now to the sectoral level. Apart from the advantage of disaggregation, we can also drop the simultaneous model, treating each sector as a price taker. However, we lack data on physical capital stocks and rentals at the sectoral level and have therefore to estimate our marginal productivity equations in their absolute rather than relative price form. This time we begin by estimating ρ from the condition that

$$\frac{\partial Y}{\partial N} = w = (1 - a)A^{\rho}\left(\frac{Y}{N}\right)^{1-\rho}$$

or, in its estimating form,

$$\log \frac{Y}{N} = -\frac{1}{1-\rho}\log\,[(1 - a)A^{\rho}] + \frac{1}{1-\rho}\log w. \tag{8}$$

We then need to estimate θ. This requires some tedious manipulations. Our remaining marginal productivity condition is for S, since we have no data on K or r. This gives

$$\frac{\partial Y}{\partial S} = z = aA^{\rho}\left(\frac{Y}{Q}\right)^{1-\rho}(1 - b)\left(\frac{Q}{S}\right)^{1-\theta}.$$

This cannot be estimated, since Q depends on θ. To eliminate Q we use Euler's theorem:[15]

$$Y = wN + \frac{\partial Y}{\partial Q}\cdot Q.$$

Now

$$\frac{\partial Y}{\partial Q} = aA^{\rho}\left(\frac{Y}{Q}\right)^{1-\rho}$$

$$\begin{aligned}\therefore\ Q &= a^{-(1/\rho)}A^{-1}(Y - wN)^{1/\rho}Y^{(\rho-1)/\rho}\\ &= a^{-(1/\rho)}A^{-1}X,\end{aligned}$$

---

[14] The comparable estimates using human capital and raw labor are 1.09 in our formulation and 1.67 where K takes the role of X_1.

[15] An alternative approach is to obtain Q from the production function relationship.

$$aQ^{\rho} = \left(\frac{Y}{A}\right)^{\rho} - (1 - a)N^{\rho}.$$

This leads to a different estimating equation from (9) which, unlike (9), uses the constant term estimate from (8) and does not use wage data. However, the estimated values of $1/(1 - \theta)$ are similar to those given by (9).

where X is defined by

$$X = (Y - wN)^{1/\rho} Y^{(\rho-1)/\rho}.$$

We now substitute for Q in the marginal productivity condition above and rearrange to obtain

$$\log \frac{X}{S} = -\frac{1}{1-\theta} \log \left[(1-b) a^{\theta/\rho} A^{\theta}\right] + \frac{1}{1-\theta} \log \left(\frac{z}{X^{\rho-1} Y^{1-\rho}}\right). \tag{9}$$

As long as A (the "efficiency parameter") is regarded as a constant across countries, equations (8) and (9) provide our basic estimating equations, the first term on the right-hand side being in each case a constant. If, however, A is considered variable, then the estimates of ρ and θ will in general be biased. Two ways were used for dealing with this. First, a *dummy variable* was included distinguishing the poorer from the richer countries. Second, an *efficiency variable* (A_i) was calculated from the whole-economy regression of Section I as follows:[16]

$$A_i = \frac{Y_i}{[aQ_i^{\rho} + (1-a)N_i^{\rho}]^{1/\rho}}.$$

We then assumed that in each sector, the ratio of the efficiency parameters of the different countries was the same as in the whole economy (as in Arrow et al. 1961). The estimating equations are got by rearranging (8) so that $1/(1-\rho)$ is the regression coefficient of $\log Y/NA$ on $\log w/A$, and by rearranging (9) so that $1/(1-\theta)$ is the regression coefficient of $\log X/SA$ on $\log z/X^{\rho-1}Y^{1-\rho}A$.

Estimates

Owing to lack of data, we do not at the sectoral level distinguish between richer and poorer countries. So for each sector we have four estimates

---

[16] The relative values are shown in table 4 (below). Since A varies between countries but is not directly observable, the model in Section I should strictly be completed by a function explaining A. In Section III we assume in a model with S' and N' that

$$\log A = \log e_0 + e_1 \log \frac{C}{N'}. \tag{10}$$

This alters the model by adding one variable and one equation and does not change the identifiability of eqq. (2'), (3'), and (4').

of the parameters, one pair using the dummy variables and the efficiency variable (A) on equations (8) and (9) and another pair using them on equations (8′) and (9′), in which S' (human capital) and N' (raw labor) appear. Since we are studying four sectors, this gives 16 estimates altogether of $[1/(1 - \rho)] - [1/(1 - \theta)]$, which must be positive if $\rho > \theta$. Of these 16 estimates, 15 are positive and seven significantly so at at least the 90 percent level (see table 2). The negative estimate is not significant. A sensitivity analysis again suggested that our t-statistics for $[1/(1 - \hat{\rho})] - [1/(1 - \hat{\theta})]$ are downward biased. Thus the capital-skill complementarity hypothesis receives strong support at the sectoral level.

As regards the direct elasticities of substitution between skill and other labor, these lie between $1(1 - \rho)$ and $1/(1 - \theta)$ and are generally not above unity.

III. Income Distribution in the Whole Economy Model

Some Stylized Facts

We now revert to the whole-economy model in its human capital version, and look at some of its implications. First we set out some stylized facts from the data and then show how the model helps to explain them. The stylized facts can be usefully summarized by stating how our key variables change when a country gets richer in terms of total capital per man (see table 3). (1) The prices of skill (z') and of raw labor (w') rise, but the price of raw labor rises faster. This reduced wage differential (z'/w') does not of itself, of course, guarantee that inequality of earnings as measured by the Gini coefficient is reduced,[17] but taken with the actual changes in relative numbers it does in fact have the effect of reducing inequality so measured.[18] This is consistent with the findings of others (Lydall 1968). It also, of course, implies a reduction in the rate of return to investing in skill. (2) The rate of return to physical capital (r) also falls absolutely, but remains roughly proportional to the rate of return on skill. (3) The ratio of physical capital relative to skill measured in physical units (K/S) rises. But the ratio of physical capital to the value of human capital ($rK/z'S'$) falls. (4) The share of raw labor in national income does not change in any systematic way, while the share of physical capital falls and that of human capital rises.

---

[17] For example, if there are two categories of labor and the poorer category grows in relative number, then the Gini coefficient can easily rise even if wage differentials are reduced.

[18] The Gini coefficient used takes no account of variations in wages within educational categories. It is simply measured as

$$\frac{1}{2}\,\frac{1}{(L_H + L_S + L_P + L_O)^2}\,\frac{1}{\bar{w}}\sum_i \sum_j L_i L_j |w_i - w_j| \qquad (i, j = H, S, P, O).$$

TABLE 2

SUBSTITUTION PARAMETERS: SECTORAL LEVEL

	USING DUMMY			USING VARIABLE A			
	Eqq. (8), (8′) $\frac{1}{1-\rho}$	Eqq. (9), (9′) $\frac{1}{1-\theta}$	$\frac{1}{1-\rho}-\frac{1}{1-\theta}$	Eqq. (8), (8′) $\frac{1}{1-\rho}$	Eqq. (9), (9′) $\frac{1}{1-\theta}$	$\frac{1}{1-\rho}-\frac{1}{1-\theta}$	NO. OF OBSERVATIONS
Mining:							
Using S, N	1.45	0.76	0.69	0.84	0.54	0.30	15
	(7.1)	(3.9)	(2.9)	(4.0)	(3.1)	(1.1)	
Using S', N'	0.93	0.84	0.09	0.48	−0.09	0.57	15
	(5.2)	(9.6)	(0.8)	(2.8)	(−0.6)	(2.5)	
Manufacturing:							
Using S, N	1.66	0.74	0.92	1.11	0.85	0.26	16
	(7.9)	(3.2)	(2.4)	(6.7)	(8.3)	(1.3)	
Using S', N'	1.16	0.91	0.25	0.63	0.40	0.23	16
	(4.2)	(8.6)	(0.4)	(3.8)	(2.6)	(1.03)	
Construction:							
Using S, N	0.90	0.25	0.65	0.89	0.66	0.23	14
	(3.5)	(1.3)	(2.0)	(3.6)	(5.3)	(0.8)	
Using S', N'	0.70	0.08	0.62	0.44	−0.44	0.88	14
	(5.4)	(0.6)	(3.4)	(2.7)	(−2.09)	(3.29)	
Electricity, gas, water:							
Using S, N	1.03	1.06	−0.03	1.09	1.01	0.08	16
	(3.6)	(39.6)	(−0.10)	(6.1)	(8.3)	(0.4)	
Using S', N'	0.63	0.23	0.40	0.69	0.22	0.47	16
	(3.2)	(1.66)	(1.66)	(4.1)	(1.4)	(2.0)	

NOTE.—Figures in parentheses are *t*-statistics (see text). The sample of countries for the *mining* sector consists of United States, Sweden, Canada, New Zealand, France, Norway, United Kingdom, Japan, Greece, Israel, Puerto Rico, Chile, Philippines, Ghana, and Thailand. The samples for the other sectors are identical except that *manufacturing* excludes Greece and includes India and Korea; *construction* excludes Greece and Ghana and includes India and Korea; *electricity and gas* includes Korea. The last column shows the number of observations for regressions using the dummy variable. In the regressions using variable *A*, Israel was excluded.

TABLE 3

CORRELATION COEFFICIENTS BETWEEN C/N' AND THE VARIABLES SHOWN

Variable	Coefficient	Variable	Coefficient
w'	.86	S'/N'	.83
r	−.79	K/N'	.96
z'	.70	K/S'	.76
z'/w'	−.45	$rK/z'S'$	−.39
r/w'	−.54	rK/Y	−.36
r/z'	−.52	$z'S'/Y$	.26
$z'/w'r$	.00	$w'N'/Y$	.06

An Explanation

Let us now see how our model relates to these facts. The estimated production function taking all countries together is

$$Y = A(0.87K^{-0.72} + 0.13S'^{-0.72})^{0.62/-0.72}N'^{0.38}. \qquad (1')$$

As can be seen, this is a Cobb-Douglas function in Q' and N', the reason being that the estimate of ρ is 0.04 (see table 1), which is not significantly different from zero ($t = 0.4$). This function therefore implies that w' rises proportionately to Y/N' and that the share of raw labor is constant.

To go further than this we need to know how growth in total capital per man is allocated between growth in physical and human capital. In our model this is determined in equation (4′), but before turning to this it is instructive to look at the implication of a simpler model in which the rates of return on human and nonhuman capital are constrained to a fixed ratio (m).[19] So

$$\frac{z'/w'}{r} = m \qquad (4'')$$

or $(z'/r) = mw'$. Given this, increased wealth per head must raise not only w' but also z'/r. And since Q is homothetic in K and S', this must raise K/S'. However, we know from table 1 that the direct elasticity of substitution between K and S' is less than unity (1/1.72), so that when K/S' rises, the ratio $rK/z'S'$ falls. In this way our production function, together with one simple behavioral constraint, predicts that the relative share of physical capital will fall with economic progress and that the shares of human capital and earnings will rise (fact 4). It also predicts that the value of human capital will rise faster than physical capital (fact 3).

As for the absolute rate of return on physical capital (r), this is bound to fall, provided A is constant, since both K/S' and K/N' are rising

---

[19] In our sample the rate of return (z'/w') is positively correlated with r, the correlation coefficient being .41. The ratio of z'/w' to r is completely uncorrelated with C/N'.

(fact 2). And by the same token the rate of return on human capital also falls (fact 1). However, as we have seen, the efficiency parameter A is not constant, and r falls with total capital per head only because the effects of changing factor proportions outweigh those of changes in efficiency.

The preceding discussion has used an exceedingly crude behavioral equation (4″) in order mainly to highlight some features of the production function. We now simulate the behavior of our key variables as C/N' changes, applying equations (1), (2′), (3′), (4′), (6′), and (7) to a country having a participation rate (N'/P) of 0.34. The supply equation estimated by two-stage least squares is

$$\log \frac{z'S'}{rK} = -0.21 + \underset{(t = 0.89)}{0.51} \log \frac{z'}{w'r} + \underset{(t = 3.4)}{0.36} \log \frac{Y}{P}. \qquad (4')$$

As can be seen, more of total capital is allocated to human capital the higher the rate of return to human capital relative to physical capital, though the estimated effect is not highly significant. Likewise, for given relative profitabilities of human and physical capital, relatively more human capital is chosen the richer the country. This reflects the psychic value attaching to human capital as such.

To see how equation (4′) brings about the rise of K/S' as C/N' rises (fact 3), we need simply combine it with (2′) to obtain

$$\log \frac{z'S'}{rK} = -3.17 + 2.22 \log w' - 1.57 \log \frac{Y}{N'}.$$

Since when Y/N' rises, w' rises at the same rate, the dependent variable rises also ($2.22 - 1.57 > 0$).

As for the fall in the absolute magnitude of r, this follows from the estimation of one further efficiency-determining equation which can be added to the model.[20]

$$\log A = 0.83 + \underset{(t = 5.84)}{0.20} \log \frac{C}{N'}. \qquad (10)$$

IV. Accounting for Variations in Output per Head

Finally, we can use our production function to analyze the sources of differences between countries in their income per head. Unfortunately a substantial part of this remains unexplained, except insofar as we choose to attribute it to efficiency differences. However, the unexplained part is a great deal less than if the role of human capital had been ignored, as columns (1–3) of table 4 show. Column 1 shows the actual income per head of each country relative to that of the United States. For column 2

---

[20] The function was actually estimated in the form $\log A = \log e_0 + e_1 \log (C/P)$, and equation (10) is obtained by substituting $N' = 0.34\ P$.

TABLE 4

SOURCES OF DIFFERENCES IN INCOME PER HEAD

Country	$\frac{(Y/P)_i}{(Y/P)_{US}}$ (1)	A_i/A_{US} 2-Factor Model (2)	A_i/A_{US} 3-Factor Model (3)	D^K (4)	D^S (5)	D^N (6)	D^A (7)	$\sum D$ (8)
Richer:								
United States...	1.00	1.00	1.00	...	...	...	...	...
Sweden	0.70	0.73	0.63	0.28	−0.47	−0.28	1.27	0.80
Canada........	0.66	0.77	0.75	0.16	0.21	0.04	0.72	1.13
N. Zealand.....	0.57	0.52	0.50	−0.39	0.05	−0.04	1.19	0.81
France	0.52	0.59	0.60	0.35	0.03	−0.13	0.85	1.10
Norway	0.50	0.58	0.60	0.16	0.26	−0.06	0.81	1.17
United Kingdom	0.49	0.59	0.61	0.52	0.03	−0.18	0.76	1.13
Japan	0.23	0.32	0.38	0.58	0.06	−0.18	0.81	1.27
Greece	0.17	0.33	0.37	0.64	0.24	−0.08	0.75	1.55
Poorer:								
Puerto Rico	0.32	0.66	0.61	0.53	0.17	0.07	0.61	1.38
Mexico	0.13	0.46	0.61	0.75	0.64	0.20	0.44	2.03
Chile	0.10	0.40	0.46	0.72	0.50	0.21	0.61	2.04
Columbia	0.09	0.29	0.33	0.72	0.38	0.12	0.73	1.95
Brazil	0.08	0.31	0.45	0.75	0.66	0.18	0.60	2.19
Turkey	0.08	0.26	0.38	0.84	0.52	−0.07	0.67	1.96
Philippines.....	0.08	0.37	0.47	0.86	0.38	0.26	0.57	2.07
Ghana	0.07	0.20	0.37	0.70	0.76	0.01	0.68	2.15
S. Korea.......	0.05	0.27	0.37	0.88	0.51	0.04	0.66	2.09
Thailand	0.04	0.16	0.23	0.84	0.43	−0.06	0.80	2.01
Kenya.........	0.03	0.18	0.28	0.85	0.74	0.13	0.74	2.46
India..........	0.03	0.11	0.22	0.85	0.74	−0.05	0.81	2.35
Uganda	0.02	0.12	0.19	0.88	0.60	−0.02	0.83	2.29

we have estimated a simple CES function using capital and raw labor only,[21] and then shown the efficiency of each *i*th country relative to that of the United States (A_i/A_{US}), where

$$A_i = \frac{Y_i}{[aK_i^\rho + (1 - a)N_i'^\rho]^{1/\rho}}.$$

The third column shows the relative efficiencies we have already used in Sections II and III. As can be seen, the differences in income per head in column 1 are greater than the efficiency differences in column 2, which in turn are greater than the efficiency differences in column 3. This pattern can be summarized by the coefficients of variation for each column which are 1.03, 0.56, and 0.40, respectively.

But although differences in factor endowments leave much unexplained, it is still useful to summarize their effects. Following Krueger (1968),

---

[21] The regression equation was

$$\log \frac{r}{w''} = 0.06 - \underset{(t\,=\,-13.75)}{1.07} \log \left(\frac{K}{N'}\right),$$

where w'' is the average wage of all workers irrespective of skill. From these parameter values, A was computed for each country.

we focus on the differences in income per head in each ith country from that in the United States. For each factor in turn we ask how much the U.S. income per head would fall if the endowment per head of that factor were reduced to its level in the ith country, all other U.S. factor endowments held constant. This fall is then expressed as a proportion of the total gap in income per head between the two countries, so as to yield a measure D_i^j indicating the relative fall due to reducing the jth factor to its level in the ith country. If the production function is written as

$$\frac{Y}{P} = Af\left[\frac{K}{P}, \frac{S'}{P}, \frac{N'}{P}\right],$$

we have, for example, that

$$D_i^K = \frac{(Y/P)_{US} - A_{US}f[(K/P)_i, (S'/P)_{US}, (N'/P)_{US}]}{(Y/P)_{US} - (Y/P)_i};$$

D_i^S and D_i^N are calculated analogously. In addition, we have

$$D_i^A = \frac{(Y/P)_{US} - A_i f[(K/P)_{US}, (S'/P)_{US}, (N'/P)_{US}]}{(Y/P)_{US} - (Y/P)_i}.$$

Clearly the statistics D do not account for income differences in the sense that they generally sum to unity. In fact, for any country less well endowed in all factors than the United States and less efficient,[22] $D_i^K + D_i^{S'} + D_i^{N'} + D_i^A \geq 1$. Nevertheless, the relative magnitudes of the D terms do provide as good an indication as any of the "importance" of different factors in explaining income difference. Physical capital is more "important" than human capital in five out of eight richer countries, and in 12 out of the 13 poorer ones.[23] This finding was already implicit

---

[22] First there is the problem of the multiplicative relation of A and f.

$$\begin{aligned} A_{US}(f_{US} - f_i) + (A_{US} - A_i)f_{US} &= A_{US}f_{US} - A_{US}f_i + A_{US}f_{US} - A_if_{US} - A_if_i + A_if_i \\ &= A_{US}f_{US} - A_if_i + (A_{US} - A_i)(f_{US} - f_i) \\ &> A_{US}f_{US} - A_if_i. \end{aligned}$$

Second, consider the interaction effects within f. To simplify, suppose f a function of two variables only, $f(K, N)$. Then, since N raises the marginal product of K,

$$f(K_{US}, N_{US}) - f(K_i, N_{US}) > f(K_{US}, N_i) - f(K_i, N_i).$$

Therefore, adding $f(K_{US}, N_{US})$ to both sides and rearranging,

$$[f(K_{US}, N_{US}) - f(K_i, N_{US})] + [f(K_{US}, N_{US}) - f(K_{US}, N_i)] > f(K_{US}, N_{US}) - f(K_i, N_i).$$

So the sum of the differences got by varying one factor at a time exceeds the overall difference $f_{US} - f_i$.

[23] Exactly the same result is obtained when we start with the lower country's endowment and calculate, for example,

$$D_i^K = \frac{A_i f[(K/P)_{US}, (S/P)_i, (N/P)_i] - (Y/P)_i}{(Y/P)_{US} - (Y/P)_i}.$$

in Section III, which showed that the proportional differences between rich and poor countries are much greater for physical capital per head than for the skill input per head measured in physical units. Since the distribution parameter in the production function is much higher for physical capital than for skill, it follows that differences in physical capital per head account for a greater proportion of differences in output per head.

This finding conflicts with that of Krueger, who offered "minimum estimates" of $D^{S'}$ higher than our own for more than half the countries covered in both our studies. In most countries her estimate exceeded one-half, on the strength of which she claimed that human capital was more important than all other factors put together (including efficiency) in explaining differences in income per head. However, she appears to have calculated the statistic

$$\frac{\sum_j w^j_{US}(L_j/N')_i}{\sum_j w^j_{US}(L_j/N')_{US}},$$

the labor force being divided into a large number of age-education-sector categories. This statistic is then presented as a minimum estimate of the income per head that would obtain in country i relative to that of the United States if country i had the U.S. endowment of physical capital per head but its own labor-force composition. If this is a correct interpretation of the statistic (in Krueger 1968, table 2, col. 2), it overestimates the amount by which output per head would in fact be reduced.[24] For assuming with Krueger that all types of labor are perfect substitutes and can be combined into a measure (E) of efficiency units per head of labor, it follows that if the elasticity of output per head with respect to E is a constant e (< 1) over the relevant range, the relevant ratios of income will be [25]

$$\left(\frac{E_i}{E_{US}}\right)^e > \frac{E_i}{E_{US}}.$$

Our conclusion about the greater relative importance of physical than human capital therefore stands.

---

[24] We are grateful to C. R. S. Dougherty for discussions on this subject; see also Dougherty (1974).

[25] The only type of measure which would definitely underestimate the share of human capital in accounting for income differences in a constant-returns production function $Y/N' = f(K/N', E)$ would be

$$\frac{(E_{US} - E_i)\cdot f_E[(K/N)_{US}, E_{US}]}{(Y/N')_{US} - (Y/N')_i}.$$

This would be an underestimate in the sense that, if added to a similar measure for the share of K/N', the sum would be less than unity. For a proof, see Krueger (1968, p. 644, note).

Appendix A

Alternative Forms of Nested CES Function

The three alternative forms of nested CES production functions using K, S, and N are:

$$Y = A[aQ^{\rho} + (1 - a)N^{\rho}]^{1/\rho}, \quad \text{where} \quad Q = [bK^{\theta} + (1 - b)S^{\theta}]^{1/\theta}; \tag{A1}$$

$$Y = A[aQ^{\rho} + (1 - a)K^{\rho}]^{1/\rho}, \quad \text{where} \quad Q = [bN^{\theta} + (1 - b)S^{\theta}]^{1/\theta}; \tag{A2}$$

$$Y = A[aQ^{\rho} + (1 - a)S^{\rho}]^{1/\rho}, \quad \text{where} \quad Q = [bK^{\theta} + (1 - b)N^{\theta}]^{1/\theta}. \tag{A3}$$

Another three functions (1′)–(3′) correspond to the permutations on K, S', and N'. We use the following criteria to distinguish between functions (1)–(3) and between functions (1′)–(3′).

a) How well does the function, when estimated using the two-stage least-square model in Section I, explain output? To answer this we use equation (A1) to compute for each country i

$$A_i = \frac{Y_i}{[aQ_i^{\rho} + (1 - a)N_i^{\rho}]^{1/\rho}},$$

and then compute the coefficient of variation $\sqrt{\sum(A_i - \bar{A})^2}/\bar{A}$. An analogous computation is also made using equations (A2) and (A3). The results are as follows.

	Eqq. (A1), (A1′)	Eqq. (A2), (A2′)	Eqq. (A3), (A3′)
Using S, N	.41	.49	.80
Using S', N'	.41	.44	.43

Functions (A1) and (A1′) perform best.

b) How stable is the estimate of the effect of X_2/X_3 on p_2/p_3 when we also let X_1 influence p_2/p_3? To answer this, we first assume X_1 separable and estimate

$$\log\frac{p_2}{p_3} = e + f\log\frac{X_2}{X_3}. \tag{A4}$$

We then assume that this is a misspecification and estimate each of the two alternative specifications, the first with X_2 separable and the second with X_3 separable:

$$\log\frac{p_2}{p_3} = e + f\log\frac{X_2}{X_3} + g\log\frac{Q^{13}}{X_3}, \quad \text{where } Q^{13} = [bX_1^{\theta} + (1 - b)X_3]^{1/\theta}; \tag{A5}$$

$$\log\frac{p_2}{p_3} = e + f\log\frac{X_2}{X_3} + g\log\frac{Q^{12}}{X_3}, \quad \text{where } Q^{12} = [bX_1^{\theta} + (1 - b)X_2]^{1/\theta}. \tag{A6}$$

To estimate functions (A5) and (A6), we used prior estimates of b and θ obtained from the corresponding fully estimated model. The resulting estimates of f are shown in table A1. For each pair (p_2, p_3) the results of the three equations (A4)–(A6) are shown on a given row. There are three rows corresponding to the three pairs (p_2, p_3) among r, z, and w, and another three rows corresponding to the three pairs among r, z', w'. As the table shows, the estimates of f are far more stable (along a row) for the pair r/z than for its two rivals, and for r/z' than for its two rivals. This provides strong support for the corresponding functions (A1) and (A1′).

TABLE A1
ESTIMATES OF f FROM EQUATIONS (A4), (A5), AND (A6)

	ADDITIONAL VARIABLE OTHER THAN X_2/X_3			
p_2/p_3	None	Q^{KS}, $Q^{KS'}$	Q^{SN}, $Q^{S'N'}$	Q^{KN}, $Q^{KN'}$
Using S, N:				
r/z	−3.45	N.A.	−1.56	−2.48
z/w	−0.28	−1.50	N.A.	−1.38
r/w	−0.81	$-\infty$	−37.07	N.A.
Using S', N':				
r/z'	−1.72	N.A.	−4.41	−1.02
z'/w'	−0.60	693.44	N.A.	−11.82
r/w'	−1.07	1.93	0.45	N.A.

NOTE.—N.A. = not applicable.

c) How sensible are the estimates of $(\theta - 1)$ and $(\rho - 1)$. and how consistent are they between rich and poor countries? This is a trickier criterion. Clearly there is no reason why $(\theta - 1)$ should be the same in both groups of countries, though one might be surprised if it differed by a multiple of 10 or if it were significantly different from zero in one group and not in the other. The same applies to $(\rho - 1)$. The estimates for equations (A1) and (A1′) are in table 1. For equations (A2) and (A2′), $(\theta - 1)$ was significant when all countries were pooled, but both very different and insignificant in rich countries and poor countries taken separately. The same applies to $(\rho - 1)$. Equations (A3) and (A3′) performed even worse; $(\theta - 1)$ was significant for all countries and for rich and poor on their own, but $(\rho - 1)$ varied wildly and was often positive—implying isoquants in Q, S space that are concave to the origin.

Appendix B

Data Sources

a) *Wages and Labor Force*

The wage data come from Psacharopoulos (1973) except for New Zealand, where they are taken from the 1966 Census. For Japan, France, and Puerto Rico only relative wages are available; absolute wages are generated from estimates of total labor income (see below). Wage data are converted into U.S. dollars using free market exchange rates where these are available, or, otherwise, official exchange rates. They are adjusted to 1963 levels by multiplying them by the ratio of monetary GDP in 1963 to monetary GDP in the year to which the data relate (these dates vary from 1959 to 1968).

The labor force data come from Psacharopoulos (1973) or Organization for Economic Cooperation and Development (1969). For most countries the sectoral breakdown is taken from the "Education by Branch of Economic Activity" tables in OECD (1969). However, this breakdown has to be generated indirectly for Canada, France, Greece, the United Kingdom, Chile, Ghana, Korea, and Puerto Rico. The method is as follows. Let A be the occupation-by-sector matrix with n rows for occupations and m columns for sectors. Let B be the occupation-by-education matrix with each element expressing the proportion of those in a given occupation having a given level of educational attainment; B consists of k rows for educational levels and n columns for occupations. Then $BA = C$ gives us a $k \times m$ matrix C, which is our estimate of the "Education by Branch of Economic Activity" table. To examine the validity of this method, we have also applied it to some of those countries for which an "Education by Branch of Economic Activity" table is given in OECD (1969). A comparison of the elements in the actual and estimated tables tends to confirm the validity of the method.

A final problem is the treatment of age. As workers get older their earnings rise, and presumably they become more productive. We therefore subdivide each educational category, such as the higher-educated, into five age classes, $L_{H1}, \ldots, L_{H5}$. We then assume that the age classes are perfect substitutes for each other, and estimate the labor input, for example, from higher-educated people as

$$L_H = \frac{\bar{W}_{H1}L_{H1} + \cdots + \bar{W}_{H5}L_{H5}}{\frac{1}{5}(W_{H1} + \cdots + \bar{W}_{H5})},$$

where $\bar{W}_{Hi}$ refers to the world average wage of the higher-educated in the ith age class (strictly the unweighted intercountry average of the country average wages). The choice of denominator is of course immaterial, but the one chosen has the effect that L_H is of the same order of magnitude as the crude number of higher-educated people in a country. The corresponding wage per unit of higher-educated labor is

$$W_H = \frac{W_{H1}\,L_{H1} + \cdots + W_H L_H}{L_H},$$

where W_{Hi} is the wage of higher-educated people in the ith age class in the country in question. The age distributions are obtained either from the United Nations (1964, 1965) or from the original samples on which the various rate-of-return studies in Psacharopoulos (1973) were based. In many cases, data are lacking on the labor force as such, and total population statistics have to be used. We assume that the proportion of workers in each age group is the same in each sector. For illiterates however, the paucity of age data for either wages or labor force made the above procedures impossible; no age adjustment was therefore made—a justifiable procedure if age-earnings profiles are roughly flat. No data on L_0 and W_0 exist for several of the richer countries. In these cases L_0 can be safely set at zero; W_0 is approximated by an average of unskilled wage rates in various sectors of the economy, as quoted in the International Labour Office *Bulletin of Labour Statistics.*

b) Output, Capital Stock, and Capital Rental

On output (value added), United Nations (1966) provides a convenient source for mining, manufacturing, and electricity, gas, and water, while the United Nations *Yearbook of National Accounts Statistics* provides the remaining data. Physical capital is measured as the simple sum of gross fixed capital formation from 1949 to 1963 inclusive, measured always at 1963 prices. Most of the data on capital formation come from various editions of the United Nations *Yearbook of National Accounts Statistics,* although for some countries national sources are also used to obtain figures for the earlier years (especially before 1953). Data given at current year prices are converted to 1963 prices using wholesale price indices from either the United Nations *Monthly Bulletin of Statistics* or from local country sources. To compute the total income of capital, GDP at factor cost is divided into employee income, income from unincorporated enterprises, and "other." Income from unincorporated enterprises is then divided between labor and capital in the ratio of employee to "other" income, and total capital income is the capital share of unincorporated income plus "other" income. Labor income is the labor share of unincorporated income plus employee income. Capital rental (per unit of capital) is total capital income divided by total capital.

References

Arrow, K. J.; Chenery, H. B.; Minhas, B. S.; and Solow, R. M. "Capital-Labor Substitution and Economic Efficiency." *Rev. Econ. and Statis.* (August 1961).

Blaug, M. "Approaches to Educational Planning." *Econ. J.* (June 1967).

Bowles, S. "Aggregation of Labour Inputs in the Economics of Growth and Planning: Experiments with a Two-Level CES Function." *J.P.E.* 78, no. 1 (January/February 1970): 68–81.

Denison, E. *Why Growth Rates Differ.* Washington: Brookings Inst., 1967.

Dougherty, C. R. S. "Estimates of Labour Aggregation Functions." *J.P.E.* 80, no. 6 (November/December 1972): 1101–1119.

———. "Factor Endowments and Per Capita Income Differences among Countries: A Note." Mimeographed. London School of Econ., 1974.

Fallon, P. R. "The Productive Roles of Workers with Different Levels of Education: An International Cross-Section Study." Ph.D. thesis, Univ. London, 1974.

Fisher, F. M. *The Identification Problem in Econometrics.* New York: McGraw-Hill, 1966.

Folger, J. K., and Nam, C. B. *Education of the American Population: A 1960 Census Monograph.* Washington: Government Printing Office, 1967.

Griliches, Z. "Capital-Skill Complementarity." *Rev. Econ. and Statis.* (November 1969).

———. "Notes on the Role of Education in Production Functions and Growth Accounting." In *Education, Income, and Human Capital,* edited by W. L. Hansen. New York: Nat. Bur. Econ. Res., 1970.

Hicks, J. R. "Elasticity of Substitution Again: Substitutes and Complements." *Oxford Econ. Papers* (November 1970).

Krueger, A. O. "Factor Endowments and Per Capita Income Differences among Countries." *Econ. J.* (September 1968).

Lydall, H. *The Structure of Earnings.* Oxford: Clarendon, 1968.

Organization for Economic Cooperation and Development. *Occupational and Educational Characteristics of the Labour Force.* Washington: OECD, 1969.

Psacharopoulos, G. *Returns to Education: An International Comparison.* Amsterdam: Elsevier, 1973.

Sato, R., and Koizumi, T. "On the Elasticities of Substitution and Complementarity." *Oxford Econ. Papers* (March 1973).

United Nations. *Demographic Yearbook.* New York, 1964, 1965.

———. *Growth in World Industry.* New York, 1966.

On Two Specifications of Asset Equilibrium in Macroeconomic Models

---

Duncan K. Foley

Stanford University

The problem of specifying asset market equilibrium conditions in a macroeconomic model is studied. Two specifications are proposed: "beginning-of-period" and "end-of-period" formulations in period models. In each case the limit of the model as the time period is shortened is rigorously calculated. It is shown that the two formulations are internally consistent, but consistent with each other only under assumptions of perfect foresight. A brief review of the liquidity preference versus loanable funds literature shows that theorists have used these models without explicitly noting the differences and inconsistencies.

1. Introduction

The purpose of this paper is to state clearly and point out some connections between three related conceptual problems in macroeconomic models: the choice of continuous versus period analysis; perfect foresight versus imperfect expectations; and stock versus flow equilibrium in the markets for assets. Although current research focuses on "disequilibrium" models, most writers have an "equilibrium" notion that lies behind their disequilibrium formulation. For this reason I will study equilibrium models.

To exhibit clearly and definitely the relations between these ideas, a simple model will be helpful. I hope it will be clear that the propositions demonstrated in this paper apply to a wide class of macroeconomic models,

I would like to thank Karen Johnson, Martin Hellwig, students and faculty in the monetary seminars at Stanford University and the University of California at Santa Barbara, and my students in monetary theory at the Massachusetts Institute of Technology for helpful comments and criticisms at various stages in the writing of this paper. Any errors remain my responsibility. This work was supported by the National Science Foundation through grant GS-39937 and in part through grant GS-40104 at the Institute for Mathematical Studies in the Social Sciences, Stanford University.

[*Journal of Political Economy*, 1975, vol. 83, no. 2]

© 1975 by The University of Chicago. All rights reserved.

both static and dynamic, and that the questions raised here must be resolved in some way in all macroeconomic models. The particular model outlined here is meant to be only an illustrative guide to make the argument more concrete.

The argument is directed to the following results: that there are two different possible specifications of asset market equilibrium in macroeconomic models, in both continuous time and period versions; that these two specifications are each internally consistent but inconsistent with each other except under the assumption of perfect foresight; that the existing macroeconomic literature does not always recognize the distinction between the two specifications and occasionally confuses them; and that certain theoretical propositions are true in one version and not in the other. One specification is "beginning-of-period" equilibrium, which corresponds in continuous time to "stock" equilibrium; the other is "end-of-period" equilibrium, which corresponds in continuous time to "flow" equilibrium.

2. A Simple Model

Time in this model can flow continuously or in periods. If periods are used, the length of the period will be called Δt. Any stock variable will be measured at the beginning of the period; flow variables will be assumed constant through a period and converted to stock variables by multiplying by Δt; and prices are assumed constant through the period and chosen so as to satisfy certain market clearing conditions.

There are two produced goods in this model: consumption goods, Q_C, which will be the numeraire, and investment goods, Q_I, assumed to be indistinguishable from existing capital. The price of investment goods in terms of consumption goods will be p_k. The two goods are produced by two factors: labor, N, which is inelastically supplied and fixed over time; and capital services, which are proportionate to the stock of physical capital, K, the sum of past output in the investment goods sector. There is no depreciation of capital. The rental (in consumption goods units per unit time) earned by one unit of capital services will be called r.

Under the assumptions of perfect mobility of factors between sectors and, for the moment, full employment of both factors, we can derive supply functions for flows of consumption and investment goods that depend on the total capital services, K, and the relative price of investment goods, p_k: $Q_C(K, p_k)$; $Q_I(K, p_k)$. The rental to capital in consumption goods units will depend only on the relative price p_k (see Rybczynski 1955; Uzawa 1961; and Foley and Sidrauski 1971, chap. 1). The rate of return to capital as an asset will have a component, r/p_k, representing current rental earnings, which will also depend only on p_k.

There are two assets in the model: money (representing liquid assets in general), M, and the stock of physical capital, K. The value of money in terms of consumption goods, p_m, is the inverse of the price level (of consumer goods in terms of money).

The expected future paths of p_k and p_m will be denoted $\hat{p}_m(t)$ and $\hat{p}_k(t)$. These paths I take as unchanged in real time when the period changes and as equal to current prices at the current period, which I will call time 0. Thus, $\hat{p}_k(0) = p_k(0)$, $\hat{p}_m(0)$, $= p_m(0)$.

The stock of physical capital is the sum or integral of past output of the investment sector, assuming no depreciation, and money is issued by the government as a "lump sum" transfer to households, D.

The laws governing the growth of nominal money and physical capital and the consumption market equilibrium condition can be written

$$M(\Delta t) - M(0) = D(0)\Delta t, \tag{1}$$

$$K(\Delta t) - K(0) = Q_I[K(0), p_k(0)]\Delta t, \tag{2}$$

$$Q_C[K(0), p_k(0)]\Delta t = C^D(\ldots)\Delta t, \tag{3}$$

where C^D is a function which reports the flow demand for consumption goods by households. Its arguments may include both stocks and flows, that is, wealth including or excluding "human" wealth, and various definitions of income. For present purposes, it does not much matter what the form or arguments of C^D actually are. These three equations are the basis of two different models in what follows.

Essentially this model has two market-determined prices: p_m, the inverse of the price level, and p_k, which in this version represents the "interest rate," since the value of p_k determines the rental component of the rate of return to capital. The pair (p_m, p_k) is constrained by the consumption market clearing equation (3). The quantity p_m may enter the demand for consumption through wealth, since changes in p_m alter the value of outside liabilities (say, high-powered money or the government debt).

There is only one market clearing relation so far to determine the pair (p_m, p_k). The additional condition naturally concerns the asset markets.

3. Two Specifications of Asset Market Equilibrium

Given their expectations about the future path of prices and rates of return, agents will choose a particular path for their consumption spending and holdings of assets. At any moment along this path for any set of current prices, we can ask the question, What configuration of assets would agents like to hold at that moment given that they are constrained to choose assets that have the same value at the given prices as the assets they actually hold? The answers to this thought experiment define a demand function for assets at each moment. The changes that agents would like to make in their asset holdings over time can be expressed in terms of these momentary demands for assets (perhaps with the addition of constraints or costs associated with certain paths).

In general, these momentary demand functions will depend on time, on total marketable wealth, and on the rates of return expected on the various assets over the near future. (Unless there is some "irreversibility" to assets, there will be no need for the agents to consider the far distant future.) The expected rates of return may be point expectations held with certainty, or distributions. In any of these circumstances, we still have a well-defined answer to the original thought experiment.

For the purposes of this paper, I will assume that the momentary demand functions for money and capital are stationary in time (that is, the agents, whenever they have the same wealth and face the same mean expected rates of return, choose the same portfolios) and I will compress any effects of uncertainty on the demands into the shapes of the functions, so that only mean expected rates of return appear explicitly as arguments. Since throughout this paper I deal only with equilibrium in a single period, if the momentary demand functions were not stationary, the basic points I am trying to establish would still hold, though the situation would be more complicated. Introducing additional moments of the expectations of prices and rates of return would also complicate matters without resolving any of the problems I want to point out.

I begin with aggregate momentary (stock) demand functions for money and capital, which will depend on total nonhuman wealth, W, and expected rates of return to the two competing assets:

$$\begin{aligned} M^d(t) = \left(\frac{1}{\hat{p}_m(t)}\right) L\biggl[\hat{W}(t), \frac{\hat{p}_m(t+\Delta t) - \hat{p}_m(t)}{\Delta t\hat{p}_m(t)}, \\ \frac{\hat{r}(t)}{\hat{p}_k(t)} + \frac{\hat{p}_k(t+\Delta t) - \hat{p}_k(t)}{\Delta t\hat{p}_k(t)}\biggr], \\ K^d(t) = \left(\frac{1}{\hat{p}_k(t)}\right) J\biggl[\hat{W}(t), \frac{\hat{p}_m(t+\Delta t) - \hat{p}_m(t)}{\Delta t\hat{p}_m(t)}, \\ \frac{\hat{r}(t)}{\hat{p}_k(t)} + \frac{\hat{p}_k(t+\Delta t) - \hat{p}_k(t)}{\Delta t\hat{p}_k(t)}\biggr]. \end{aligned}$$

Notice that the functions L and J report the real value of money and capital in terms of consumption good units that the agents want to hold, given their expected wealth and expected rates of return.

The rate of return to holding money is the rate of expected change in the value of money per unit time. The rate of return to holding capital is the rental rate of return plus the expected change in the value of capital per unit time. Nonhuman wealth at time t has to be the value of money and capital in existence at time t. For any values of $\hat{p}_m$, $\hat{p}_k$, it must be true that

$$L + J = \hat{p}_m(t)M^d(t) + \hat{p}_k(t)K^d(t) = \hat{p}_m(t)\hat{M}^s(t) + \hat{p}_k(t)\hat{K}^s(t) = \hat{W}(t), \tag{A}$$

where $\hat{M}^s$ and $\hat{K}^s$ are expected holdings of money and capital at time t. Agents must plan to hold their wealth in the form of either money or capital at each instant. This balance sheet relation is prior to and separate from any Walras's law–type relation we may derive concerning flow demands. I will call it the "balance sheet constraint."

The question now arises as to how equilibrium in the markets for money and capital ought to be characterized. I believe that there are two distinct characterizations in use by theorists which have quite different theoretical consequences. Theorists are not always explicit about which version they are using, nor are they invariably careful in coping with the difficulties each version presents. I will consider each in turn.

A. End-of-Period Equilibrium

In a period model, it seems reasonable that agents will look ahead in making their trading plans and try to anticipate their desired position as of the end of the period in formulating their market demands. In this view, the operative demand for capital (or money) will be the amount of capital (or money) the agents think, at time 0, that they will want to hold at the beginning of the next period, that is, at time Δt. The appropriate supply of capital will be the sum of the existing stock of capital and the additions to it cumulated over the period. This view leads to asset equilibrium conditions:

$$\begin{aligned} K(0) + Q_I[K(0), p_k(0)]\Delta t = \frac{1}{\hat{p}_k(\Delta t)} J\Bigg[& \hat{W}(\Delta t), \frac{\hat{p}_m(2\Delta t) - \hat{p}_m(\Delta t)}{\Delta t \hat{p}_m(\Delta t)}, \\ & \frac{\hat{r}(\Delta t)}{\hat{p}_k(\Delta t)} + \frac{\hat{p}_k(2\Delta t) - \hat{p}_k(\Delta t)}{\Delta t \hat{p}_k(\Delta t)}\Bigg] = K^D(\Delta t), \end{aligned} \tag{4}$$

$$\begin{aligned} M(0) + D(0)\Delta t = \frac{1}{\hat{p}_m(\Delta t)} L\Bigg[& \hat{W}(\Delta t), \frac{\hat{p}_m(2\Delta t) - \hat{p}_m(\Delta t)}{\Delta t \hat{p}_m(\Delta t)}, \\ & \frac{\hat{r}(\Delta t)}{\hat{p}_k(\Delta t)} + \frac{\hat{p}_k(2\Delta t) - \hat{p}_k(\Delta t)}{\Delta t \hat{p}_k(\Delta t)}\Bigg] = M^D(\Delta t). \end{aligned} \tag{5}$$

These equations determine $p_m(0)$ and $p_k(0)$ given expectations. Notice that the appropriate rates of return on money and capital are those expected for next period, not for this period, since the agent is trying to adjust its portfolio to be correct at the beginning of next period.

The quantity $\hat{W}$ is wealth expected at the beginning of the next period. It clearly must be related to planned saving over the period. The difference between current wealth and wealth expected at the beginning of the next period has three components: the value of saving at current-period market prices; the capital gain or loss expected on beginning-of-period capital and money due to changes in p_m and p_k at the end of the period; and the capital gain on the money and capital accumulated during the period due to changes in p_m and p_k at the end of the period. Since income, Y, is equal to the value of output plus the value of government transfers, we can write

$$\begin{aligned}\hat{W}(\Delta t) - W(0) &= [Q_C + p_k(0)Q_I + p_m(0)D - C^D]\Delta t \\ &\quad + [\hat{p}_k(\Delta t) - p_k(0)]K(0) \\ &\quad + [\hat{p}_k(\Delta t) - p_k(0)][K^D(\Delta t) - K(0)] \\ &\quad + [\hat{p}_m(\Delta t) - p_m(0)]M(0) \\ &\quad + [\hat{p}_m(\Delta t) - p_m(0)][M^D(\Delta t) - M(0)] \\ &= [Q_C - C^D]\Delta t \\ &\quad + \{p_k(0)Q_I\Delta t + [\hat{p}_k(\Delta t) - p_k(0)]K^D(\Delta t)\} \\ &\quad + \{p_m(0)D\Delta t + [\hat{p}_m(\Delta t) - p_m(0)]M^D(\Delta t)\}.\end{aligned} \tag{6}$$

This expression leads directly to the demonstration that "Walras's law" holds for the three markets in consumption, money, and capital in the sense that the sum of the values of excess demand at any (p_m, p_k) pair in three markets will be zero, since, using equation (A), and the facts that

$$\hat{p}_k(\Delta t)K^D(\Delta t) + \hat{p}_m(\Delta t)M^D(\Delta t) = \hat{W}(\Delta t)$$

and

$$p_m(0)M(0) + p_k(0)K(0) = W(0),$$

$$\begin{aligned}(Q_C - C^D)\Delta t &= p_kK^D(\Delta t) + p_mM^D(\Delta t) - W(0) - p_kQ_I\Delta t - p_mD\Delta t \\ &= p_k[K^D(\Delta t) - K(0) - Q_I\Delta t] \\ &\quad + p_m[M^D(\Delta t) - M(0) - D\Delta t],\end{aligned} \tag{7}$$

which is just the "Walras's law" statement. If households plan their savings and consumption behavior using the same expected paths for prices, the sum of excess demands in the three markets will be equal.

In this end-of-period equilibrium model, one of the three market clearing equations is redundant, and any one of (3), (4), or (5) could be eliminated, leaving two market equilibrium conditions to determine the two prices p_m and p_k. We could even, for instance, eliminate the consumption market clearing relation and operate only with the markets for capital and money.

B. Beginning-of-Period Equilibrium

There is an alternative specification of asset market equilibrium. We may suppose that prices and rates of return adjust so as to make wealth holders content at the beginning of the period with current asset supplies.

This idea leads to the following asset market equilibrium conditions:

$$K(0) = \frac{1}{p_k(0)} J\left[p_k(0)K(0) + p_m(0)M(0), \frac{\hat{p}_m(\Delta t) - p_m(0)}{\Delta t p_m(0)}, \frac{r(0)}{p_k(0)} + \frac{\hat{p}_k(\Delta t) - p_k(0)}{\Delta t p_k(0)}\right], \tag{4b}$$

$$M(0) = \frac{1}{p_m(0)} L\left[p_k(0)K(0) + p_m(0)M(0), \frac{\hat{p}_m(\Delta t) - p_m(0)}{\Delta t p_m(0)}, \frac{r(0)}{p_k(0)} + \frac{\hat{p}_k(\Delta t) - p_k(0)}{\Delta t p_k(0)}\right]. \tag{5b}$$

Either one of these conditions is implied by the other through the balance sheet constraint, since $L + J \equiv p_m(0)M(0) + p_k(0)K(0)$. However, "Walras's law" does *not* hold between the three conditions (3), (4b), and (5b). It is easy to find (p_m, p_k) pairs that satisfy (4b) and necessarily (5b) without satisfying (3). The reason for this is that the balance sheet constraint forces the sum of the excess demands in the assets markets in this beginning-of-period sense to be zero, regardless of excess demand in the market for consumption goods. Whereas in the end-of-period equilibrium case we can eliminate any one of the three market clearing conditions, including the consumption market, in the beginning-of-period case we can eliminate any *asset* market clearing condition but we must retain the consumption market clearing condition, which always contains separate information.

C. The "Scenarios"

What are the economic pictures that lie behind these two mathematical formulations? In the end-of-period equilibrium, demands and supplies are offered as of the end of the period. Agents can offer to sell capital, for instance, which does not exist at the trading moment but which they plan to produce during the period. Contracts are made for labor and capital services and consumption during the period and asset deliveries at the end.

In the beginning-of-period equilibrium, trading in and delivery of assets are assumed to take place at the same time, that is, the beginning of the period. Within-period consumption is contracted for, but within-period production of new capital is done on a kind of speculation, taking the current-period price of capital as a guide.

It is somewhat surprising from an economic (though not from a mathematical) point of view that this very minor difference in conceptualization of equilibrium trading should give rise to sharply different models.

4. Continuous and Period Analysis

The arguments of this section are based on a methodological precept concerning macroeconomic period models: *No substantive prediction or explanation in a well-defined macroeconomic period model should depend on the real time length of the period.*[1]

In reality, economic transactions do not take place continuously. Purchases of final goods, capital account transactions, wage payments, and dividend payments all take place at discrete times. When the rhythm of these transactions is very rapid, however, and when transactions overlap for different economic agents, it is tempting to model transactions as taking place continuously. We must remember, however, that this procedure is only an approximation to the real situation.

Although individual transactions are discrete, they are not synchronized as among various agents. Therefore, period models, which assume that all transactions of a certain class occur in the same, synchronized rhythm, are also an approximation to reality. "First, it must be admitted that the period analysis is highly artificial, since while people may take decisions discontinuously, not all people take decisions at the same time" (Hahn 1955, p. 64). The proper method of modeling transactions would presumably be to allow time to flow continuously and to view transactions as discrete events of zero duration, like the "arrivals" in a Poisson process.

There might be some circumstances in which a "natural period" for economic activity can be identified. In a largely agricultural, single-crop economy, almost all plans for a single growing season might be made at the same moment. But in modern industrial economies there seems to be no presumption that plans are all made at the same calendar time. Some writers have argued that the period in the period model is economically determined, say, as the planning period of the agents. This approach as a justification for a macroeconomic-period model is fallacious because it does not speak to the question of synchronization. If all agents have the same planning period but their initial planning dates are distributed uniformly over calendar time, a continuous model would be more plausible.

---

[1] See May (1970) for a lucid comment on Patinkin's model, which introduced this principle as a basis for a rigorous criticism of macroeconomic models.

A theorist using a period model must either establish a natural period in which decisions of many different agents are synchronized or accept the position that a period model is, like the continuous model, an approximation to reality, in which case the outcomes of the macroeconomic model should not depend in any important way on the period used. A period model should give the same results (allowing for second-order inaccuracies due to coarseness of approximation) when the period is halved or doubled.

Period models are attractive because they are easy to conceptualize. This advantage is bought at a price, however, which is the danger that hidden assumptions may enter the model under the cover of the fixed period length. If the results of a period model do not depend in any important way on the period, the model can be formulated as a continuous model. The method used to accomplish this is to retain the length of the period as an explicit variable in the mathematical formulation of a period model and to make sure that it is possible to find meaningful limiting forms of the equations as the period goes to zero. (This is the procedure May [1970] uses to criticize Patinkin's model.) In my view, this procedure should be routinely applied as a test that any period model is consistent and well formed where no particular calendar time is specified as the natural period.

The remainder of this section will be devoted to applying this test to the two versions of asset equilibrium introduced in section 3.

First, there is no difficulty with equations (1)–(3) in going to the limiting case where Δt becomes very small, since by inspection they have the limiting forms

$$\frac{dM}{dt} = D(0), \tag{1*}$$

$$\frac{dK}{dt} = Q_I[K(0), p_k(0)], \tag{2*}$$

$$Q_C[K(0), p_k(0)] = C^D(\dots). \tag{3*}$$

A. End-of-Period Equilibrium and "Flow" Equilibrium

The end-of-period equilibrium conditions as written in equations (4) and (5) do not pass the test proposed above, since the equilibrium prices of money and capital will change drastically as the period becomes shorter and will in fact not exist at all for very small Δt. To see this, rewrite (4) in the following form:

$$K^D(\Delta t) - K^D(0) + K^D(0) - K(0) = Q_I[K(0), p_k(0)]\Delta t.$$

We have added and subtracted $K^D(0)$, desired holdings of capital at the beginning of the current period at current period prices. The quantity $K^D(0)$ may or may not be equal to $K(0)$, since current period prices, according to (4) and (5), are chosen to equalize prospective end-of-period demand and supplies and may not equalize beginning-of-period demands and supplies.

If we now divide by Δt, we get

$$\left[\frac{K^D(\Delta t) - K^D(0)}{\Delta t}\right] + \left[\frac{K^D(0) - K(0)}{\Delta t}\right] = Q_I[K(0), p_k(0)].$$

If we let Δt shrink, the first term on the left-hand side has a finite limit, dK^D/dt, which will be derivable from the function J. But the second term has no finite limit, so that the equilibrium condition is revealed to be ill-formed.

The problem can be seen intuitively. The household, in planning its end-of-period stock of capital, has two goals in mind: first, to correct its current portfolio imbalances and, second, to allow for its planned saving over the period. The hidden assumption is that the household adjusts its portfolio imbalance completely within the period.[2] If the period becomes short, the size of this adjustment term, since it stays constant in absolute terms, overwhelms the flow of investment and saving.

One solution to this problem is to make the speed of adjustment of current portfolio imbalance independent of the model period. Introduce a new parameter, k, measured in percent per unit time, the speed of adjustment of existing portfolio imbalance. Then the market clearing conditions in capital and money markets will be

$$K^D(\Delta t) - K^D(0) + k\Delta t[K^D(0) - K(0)] = Q_I[K(0), p_k(0)]\Delta t, \tag{4a}$$

$$M^D(\Delta t) - M^D(0) + k\Delta t[M^D(0) - M(0)] = D(0)\Delta t. \tag{5a}$$

These revised flow demands (the left-hand sides of [4a] and [5a]) also imply that the sum of the values of excess demands at current prices in the three markets will be zero. To see this, note that that balance sheet constraint implies that

$$p_kK^D(0) + p_mM^D(0) = p_kK(0) + p_mM(0) = W(0),$$

which implies that

$$k\Delta tp_k[K^D(0) - K(0)] + k\Delta tp_m[M^D(0) - M(0)] = 0,$$

so that, by (7),

$$\begin{aligned} &p_k[K^D(\Delta t) - K^D(0)] + k\Delta tp_k[K^D(0) - K(0)] - p_kQ_I[K(0), p_k]\Delta t \\ &\quad + p_m[M^D(\Delta t) - M^D(0)] + k\Delta tp_m[M^D(0) - M(0)] - p_mD(0)\Delta t \\ &\qquad = p_kK^D(\Delta t) - p_kK^D(0) - p_kQ_I[K(0), p_k(0)]\Delta t \qquad (8) \\ &\qquad\quad + p_mM^D(\Delta t) - p_mM^D(0) - p_mD(0)\Delta t \\ &\qquad = (Q_C - C^D)\Delta t. \end{aligned}$$

---

[2] This point and the discussion immediately following were first brought to my attention in a private communication from Polly Reynolds Allen.

Introduction of the parameter k also allows us to derive a limiting continuous form of the model. Dividing both sides of (4a) and (5a) by Δt, we get

$$\frac{K^D(\Delta t) - K^D(0)}{\Delta t} + k[K^D(0) - K(0)] = Q_I[K(0), p_k(0)],$$

$$\frac{M^D(\Delta t) - M^D(0)}{\Delta t} + k[M^D(0) - M(0)] = D(0).$$

We can now take the limits of these equilibrium conditions as Δt goes to zero.[3]

---

[3] We can approximate $\hat{p}_m(2\Delta t)$ as

$$\begin{aligned}\hat{p}_m(2\Delta t) &\approx \hat{p}_m(\Delta t) + \left.\frac{d\hat{p}_m}{dt}\right|_{\Delta t}\Delta t\\ &\approx p_m(0) + \left.\frac{d\hat{p}_m}{dt}\right|_0 \Delta t + \left(\left.\frac{d\hat{p}_m}{dt}\right|_0 + \left.\frac{d^2\hat{p}_m}{dt^2}\right|_0 \Delta t\right)\Delta t\\ &= p_m(0) + 2\left.\frac{d\hat{p}_m}{dt}\right|_0 \Delta t + \left.\frac{d^2\hat{p}_m}{dt^2}\right|_0 \Delta t^2,\end{aligned}$$

using

$$p_m(\Delta t) \approx p_m(0) + \left.\frac{d\hat{p}_m}{dt}\right|_0 \Delta t$$

and

$$\left.\frac{d\hat{p}_m}{dt}\right|_{\Delta t} \approx \left.\frac{d\hat{p}_m}{dt}\right|_0 + \left.\frac{d^2\hat{p}_m}{dt^2}\right|_0 \Delta t.$$

Therefore,

$$\begin{aligned}&\lim_{\Delta t\to 0}\frac{J_2}{\Delta t}\left[\frac{\hat{p}_m(2\Delta t) - \hat{p}_m(\Delta t)}{\Delta t\hat{p}_m(\Delta t)} - \frac{\hat{p}_m(\Delta t) - p_m(0)}{\Delta t p_m(0)}\right]\\ &= \lim_{\Delta t\to 0}\frac{J_2}{(\Delta t)^3 p_m(0)\hat{p}_m(\Delta t)}[\hat{p}_m(2\Delta t)p_m(0)\ \Delta t - \hat{p}_m(\Delta t)\ \Delta t p_m(0)\\ &\qquad - \hat{p}_m(\Delta t)\hat{p}_m(\Delta t)\ \Delta t + \hat{p}_m(\Delta t)\ \Delta t p_m(0)]\\ &= \lim_{\Delta t\to 0}\frac{J_2}{(\Delta t)^2 p_m(0)\hat{p}_m(\Delta t)}\left\{p_m(0)\left[p_m(0) + 2\left.\frac{d\hat{p}_m}{dt}\right|_0 \Delta t + \left.\frac{d^2\hat{p}_m}{dt^2}\right|_0 (\Delta t)^2\right]\right.\\ &\qquad \left. - \left[p_m(0) + \left.\frac{d\hat{p}_m}{dt}\right|_0 \Delta t\right]\left[p_m(0) + \left.\frac{d\hat{p}_m}{dt}\right|_0 \Delta t\right]\right\}\\ &= \lim_{\Delta t\to 0}\frac{J_2}{(\Delta t)^2 p_m(0)\hat{p}_m(\Delta t)}\left[p_m(0)p_m(0) + 2p_m(0)\left.\frac{d\hat{p}_m}{dt}\right|_0 \Delta t\right.\\ &\qquad \left. + p_m(0)\left.\frac{d^2\hat{p}_m}{dt^2}\right|_0 (\Delta t)^2 - p_m(0)p_m(0) - 2p_m(0)\left.\frac{d\hat{p}_m}{dt}\right|_0 \Delta t - \left(\left.\frac{d\hat{p}_m}{dt}\right|_0\right)^2 (\Delta t)^2\right]\\ &= \lim_{\Delta t\to 0}\frac{J_2}{(\Delta t)^2 p_m(0)\hat{p}_m(\Delta t)}\left[\left.\frac{d^2\hat{p}_m}{dt^2}\right|_0 p_m(0)(\Delta t)^2 - \left(\left.\frac{d\hat{p}_m}{dt}\right|_0\right)^2 (\Delta t)^2\right]\\ &= \lim_{\Delta t\to 0}\frac{J_2}{p_m(0)\hat{p}_m(\Delta t)}\left[\left.\frac{d^2\hat{p}_m}{dt^2}\right|_0 p_m(0) - \left(\left.\frac{d\hat{p}_m}{dt}\right|_0\right)^2\right]\\ &= J_2\left.\frac{d\pi_m}{dt}\right|_0,\end{aligned}$$

where $\pi_m = (d\hat{p}_m/dt)/p_m$. The other terms in the limit are derived in a similar fashion.

The continuous time version of the end-of-period or flow equilibrium asset market clearing conditions are

$$
\begin{aligned}
-\pi_k K^D(0) + \frac{1}{p_k}\left\{\frac{d\hat{W}}{dt} J_1 + \frac{d\pi_m}{dt} J_2 + \left[\frac{d\widehat{(r/p_k)}}{dt} + \frac{d\pi_k}{dt}\right] J_3\right\} \\
+ k[K^D(0) - K(0)] \qquad (4a^*) \\
= Q_I[K(0), p_k(0)],
\end{aligned}
$$

$$
\begin{aligned}
-\pi_m M^D(0) + \frac{1}{p_m}\left\{\frac{d\hat{W}}{dt} L_1 + \frac{d\pi_m}{dt} L_2 + \left[\frac{d\widehat{(r/p_k)}}{dt} + \frac{d\pi_k}{dt}\right] L_3\right\} \\
+ k[M^D(0) - M(0)] \qquad (5a^*) \\
= D(0),
\end{aligned}
$$

where $\pi_k = (d\hat{p}_k/dt)/p_k$ and $\pi_m = (d\hat{p}_m/dt)/p_m$ are the expected rates of change in capital and money prices in terms of consumption goods.

It seems natural to call (4a*) and (5a*) "flow" equilibrium conditions, since they equate an instantaneous flow demand for an asset with an instantaneous flow supply. It is clear that in continuous time the flow equilibrium condition corresponds to end-of-period equilibrium in a period model.

The first term represents the fact that expected price changes will satisfy some of the increased desire to hold the asset. The second term shows the flow demand for the asset that arises from current planned saving. The third and fourth terms represent revision of portfolio balance in response to expected changes in rates of return to the two assets. Notice that these terms involve *second-order expectations*, that is, expectations about the second time derivative of prices and about the rate of change (the first time derivative) of rates of return. (Since there are no bonds in this model, the only term that reflects this is $[d/dt][r/p_k]$. If there were bonds there would be terms involving $d\hat{i}/dt$, the expected change of the interest rate.)

The closed continuous flow equilibrium model consists of equations (3*), (4a*), and (5a*). Walras's law holds in these three markets taken together, so that any one is redundant and we are left with two equilibrium conditions to determine the two market variables p_m and p_k.

In order to achieve a consistent formulation of this model, we have been forced, however, to introduce an additional arbitrary parameter, k, the speed at which wealth holders adjust their portfolios. Ideally this speed should be derived from an underlying theory of the real costs to portfolio adjustment and should be an economically determined variable in the system.[4]

---

[4] This point arose in a conversation with Martin Hellwig.

B. Beginning-of-period Equilibrium and "Stock" Equilibrium

The continuous time version of beginning-of-period equilibrium is easy to derive. As Δt becomes small, the expected rates of return approach their instantaneous derivative values and nothing else changes. The resulting continuous time asset market equilibrium conditions are

$$K(0) = \frac{1}{p_k(0)} J\left[p_k(0)K(0) + p_m(0)M(0), \pi_m, \frac{r(0)}{p_k(0)} + \pi_k\right], \tag{4b*}$$

$$M(0) = \frac{1}{p_m(0)} L\left[p_k(0)K(0) + p_m(0)M(0), \pi_m, \frac{r(0)}{p_k(0)} + \pi_k\right], \tag{5b*}$$

where $\pi_m = (d\hat{p}_m/dt)/p_m$ and $\pi_k = (d\hat{p}_k/dt)/p_k$ are as before.

It seems natural to call (4b*) and (5b*) "stock" equilibrium conditions, since they equate an instantaneous demand to hold a stock of an asset with an instantaneous stock supply. It is clear that in continuous time the stock equilibrium conditions correspond to beginning-of-period equilibrium in a period model.

In the continuous model (3*), (4b*), and (5b*), either one of the asset market clearing conditions can be eliminated, but the consumption market clearing condition must be retained.

5. Stock and Flow Equilibrium and Perfect Foresight

Though we seem at this point to be faced with the necessity of choosing between the beginning-of-period or stock equilibrium model and the end-of-period or flow equilibrium model of assets markets, there is still one glimmer of hope that we can avoid making the choice. It might be that both hold simultaneously, that they are not inconsistent but in fact are only two different ways of writing the same condition.

There is an additional assumption that makes the two equilibrium conditions consistent, the assumption that the paths of expected prices ($\hat{p}_m$, $\hat{p}_k$ in the model I have been using as an example) coincide with the paths of actual prices. I will call this assumption the "perfect-foresight" assumption. Perfect foresight may hold over all time or only over certain periods of time.[5]

That perfect foresight makes the two equilibrium conditions consistent can be seen intuitively by considering the period model (4a) and (5a), where the period is chosen to make $k\Delta t = 1$. If the beginning-of-period equilibrium condition holds in both the first and second period, agents' portfolios at the end of the first period will be exactly the ones they would have chosen if they had expected prices $p_m(\Delta t)$ and $p_k(\Delta t)$. But on a perfect-foresight part of the path, these will be the values of $\hat{p}_m(\Delta t)$ and $\hat{p}_k(\Delta t)$, so the end-of-period equilibrium condition will also be fulfilled in the first period.

---

[5] Notice that "perfect foresight" in this sense is a property only of the mean expected price. In more complex models where expectations are represented as probability distributions, this notion of "perfect foresight" corresponds to the assumption that the mean of that distribution is correct. It does not correspond to the idea that as the period becomes short the variance of that distribution goes to zero.

It may at first seem that in continuous time, perfect foresight must hold because $\hat{p}_m(0) = p_m(0)$ and $\hat{p}_k(0) = p_k(0)$, so that as Δt becomes small the expected level of prices comes close to the actual level. In fact, the problem is transferred into the first derivative of price expectations in continuous time: although agents always know what price rules at any moment, they may be mistaken about the current rate of change in the price. To require agents in continuous time to have correct notions of the rate of change of prices is mathematically equivalent to requiring that they have correct expectations of the next period's level of prices in period analysis. Any objection to the assumption of perfect foresight in a period model applies with equal force to the assumption of correct notions of first derivatives of prices in a continuous model.

The precise relations between stock and flow equilibrium and perfect foresight can be summed up in the following propositions:

Proposition I: The assumption of stock equilibrium and the assumption of perfect foresight together imply that the flow equilibrium conditions will be met.

Proof: I will argue this in the continuous time model. The stock equilibrium condition, which holds at every instant, is

$$K = \frac{1}{p_k} J\left(p_k K + p_m M, \pi_m, \frac{r}{p_k} + \pi_k\right). \tag{4b*}$$

Differentiate this with respect to time to get

$$\frac{dK}{dt} = -\frac{dp_k}{dt}(1/p_k^2)J + \frac{1}{p_k}\left\{J_1\left(\frac{dp_k}{dt}K + p_k\frac{dK}{dt} + \frac{dp_m}{dt}M + p_m\frac{dM}{dt}\right)\right.$$
$$\left. + J_2\frac{d\pi_m}{dt} + J_3\left[\frac{d(r/p_k)}{dt} + \frac{d\pi_k}{dt}\right]\right\}.$$

This can be rewritten

$$Q_I(K, p_k) = -\left(\frac{dp_k}{dt}\right)(1/p_k^2)J$$
$$+ \frac{1}{p_k}\left\{\frac{dW}{dt}J_1 + J_2\frac{d\pi_m}{dt} + J_3\left[\frac{d(r/p_k)}{dt} + \frac{d\pi_k}{dt}\right]\right\},$$

which is exactly the same as (4a*) if expectations are correct (which implies $\pi_k = [dp_k/dt]/p_k$ and $\pi_m = [dp_m/dt]/p_m$), since $K^D(0) - K(0) = 0$ by the hypothesis of stock equilibrium.

Proposition II: The assumption of stock equilibrium and the assumption of flow equilibrium, if they hold for arbitrary asset demand functions, imply perfect foresight in p_m, p_k, and r/p_k.

Proof: Differentiating the stock equilibrium conditions as in the proof of Proposition I, letting $K^D(0) - K(0) = M^D(0) - M(0) = 0$ in the flow equilibrium conditions and equating the two equilibrium conditions, yields

$$\left(\frac{dp_k}{dt}\frac{1}{p_k} - \pi_k\right)J + \left(\frac{d\hat{W}}{dt} - \frac{dW}{dt}\right)J_1 + \left[\frac{d(\widehat{r/p_k})}{dt} - \frac{d(r/p_k)}{dt}\right]J_3 = 0,$$

$$\left(\frac{dp_m}{dt}\frac{1}{p_m} - \pi_m\right)L + \left(\frac{d\hat{W}}{dt} - \frac{dW}{dt}\right)L_1 + \left[\frac{d(\widehat{r/p_k})}{dt} - \frac{d(r/p_k)}{dt}\right]L_3 = 0.$$

From equation (6) we can show that

$$\frac{d\hat{W}}{dt} - \frac{dW}{dt} = \left(\frac{d\hat{p}_k}{dt} - \frac{dp_k}{dt}\right)K + \left(\frac{d\hat{p}_m}{dt} - \frac{dp_m}{dt}\right)M$$

when $K^D(0) = K(0)$ and $M^D(0) = M(0)$. In the capital market this yields, after a little manipulation,

$$-\left(\pi_k - \frac{dp_k}{dt}\frac{1}{p_k}\right)JL_1 + \left(\pi_m - \frac{dp_m}{dt}\frac{1}{p_m}\right)LJ_1 = -J_3\left[\frac{d(\widehat{r/p_k})}{dt} - \frac{d(r/p_k)}{dt}\right].$$

This equation can hold for arbitrary JL_1, LJ_1, and J_3 only if

$$\pi_k = \frac{dp_k}{dt}\frac{1}{p_k}, \qquad \pi_m = \frac{dp_m}{dt}\frac{1}{p_m}, \qquad \text{and} \qquad \frac{d(\widehat{r/p_k})}{dt} = \frac{d(r/p_k)}{dt},$$

so that expectations are in fact correct.

Clearly there will be some special cases where a particular set of asset demand functions and particular nonperfect foresight paths for expectations are compatible with both stock and flow equilibrium conditions holding simultaneously, but, in general, stock and flow equilibrium together imply perfect foresight.

It is not true that flow equilibrium plus perfect foresight will imply stock equilibrium in any finite time, because the general flow equilibrium condition assumes only a gradual adjustment of an initial stock disequilibrium.

However, the theoretical moral of these propositions is that we can avoid choosing between the stock equilibrium model and the flow equilibrium model only if we are willing to assume perfect foresight (as is often done in microeconomic capital theory models).

The decisive objection, in my opinion, to the assumption of perfect foresight in a macroeconomic model is that macroeconomics is in large part directed toward understanding the effects of discretionary government policy changes on the economy and perfect foresight rules out all government policy changes except those that are fully anticipated and announced far ahead of the time they are taken. It also seems odd to assume that the private economy predicts the course of asset prices correctly when the history of individual firms, the distribution of wealth, and asset prices themselves are written largely in terms of unanticipated movements of asset prices.

While it is logically satisfying to assume perfect foresight and in that way reconcile the two possible asset market equilibrium models, the practical need to treat imperfect anticipations forces us to choose in many practical applications between stock and flow equilibrium models.

6. Stock versus Flow Equilibrium

Given that a choice must be made between the stock and flow equilibrium conditions in assets markets, on what grounds can it be made?

In the continuous time versions of the two models, we can show that in a formal sense the stock equilibrium conditions are the limit of the flow equilibrium conditions as the parameter k becomes infinite. Rewriting (4a*), for instance, we get

$$K^D(0) - K(0) = \frac{1}{k}\Bigg(Q_I[K(0), p_k(0)] + \pi_k K^D(0) - \frac{1}{p_k}\left\{\frac{d\hat{W}}{dt}J_1 + \frac{d\pi_m}{dt}J_2 + \left[\frac{d\widehat{(r/p_k)}}{dt} + \frac{d\pi_k}{dt}\right]J_3\right\}\Bigg).$$

Since the expression in large parentheses is bounded, as $k \to \infty$ this equation will read $K^D(0) - K(0) = 0$, which is the same as (4b*).

However, it is not clear that the same equivalence holds mathematically in a period model. If we keep Δt constant, the modified "end-of-period" equilibrium conditions (4a) and (5a) are difficult to interpret if $k\Delta t > 1$, since they imply an overadjustment of initial portfolio imbalance within the period. (If $k\Delta t = 2$, for instance, wealth holders would be adding to their demand for end-of-period assets twice the amount of their initial portfolio imbalance.)

In light of the precept discussed in section 4, the fact that this limiting equivalence of stock and flow equilibrium holds only in the continuous time versions of the models should make us rather cautious in interpreting the equivalence result. We may be confronted here with a difference in "visions" of the economic process: one, stock equilibrium, that conceives of the whole of existing assets being at least potentially on the market at any moment; and another, flow equilibrium, that conceives of only a tiny fraction of the existing assets, of the same order of magnitude as additions to stocks of assets, being on the market at any moment.

The equivalence does suggest, however, that there will be very little difference between the predictions of the models when speeds of adjustment of portfolio imbalance are high or, equivalently, when costs of portfolio adjustment are low. Asset markets are in fact among the best organized of markets; information about prices of many (especially financial) assets is disseminated widely and rapidly, and the great bulk of the total wealth in industrialized capitalist economies is held in very large portfolios for which fixed transaction costs will be negligible in relation to portfolio shifts. These observations suggest that the vision of stock equilibrium may be a good approximation to the real situation. Empirical evidence of large transaction costs would, of course, upset this conclusion.

In passing, we might note some other differences between the two models, which have more pedagogic than research significance. The stock equilibrium conditions are uniformly easier to manipulate than the flow conditions. Comparative statics in the flow model require assumptions about the second derivatives of asset demand functions, while the same exercises with stock equilibrium require only assumptions about first derivatives. The much used (by stock theorists) and much criticized (by flow theorists) "*IS-LM*" apparatus can be rigorously justified in the stock model, since asset market equilibrium is separate from the consumption goods market but no sense in the flow equilibrium model, where Walras's law operates in the three markets together.

My own feeling as I write this is that the flow equilibrium model complicates our view of reality without explaining many of the significant phenomena left unexplained by the stock equilibrium model. This is probably a minority view among workers in the field.

7. Confusions over the Distinction

A word may be in order about how easy it is for these issues to become confused in scholarly debate. The choice between stock and flow equilibrium is often suppressed to the level of the "model in the back of one's mind" rather than brought out explicitly as part of the specification. Even in the back of people's minds, the issue may be confused by implicit reference to the perfect-foresight case where the two models become one, especially since the perfect-foresight assumption is the natural intertemporal generalization of the assumption in microeconomic theory that agents have perfect information about prices in all markets. A theorist consulting his or her "economic intuition," that is, the basic micro model, may come up with a model in which incompatible fragments of a stock equilibrium model and a flow equilibrium model are combined.

The confusion problem is compounded by the fact that, because the stock excess demands are part of the flow demands, most ordinary theoretical statements (such as a fall in the interest rate, other things equal, increases excess demand for money) will be true in both specifications. The disagreement will show up only in discussions of Walras's law or such controversies as: does excess supply in the (consumption) goods market imply an offsetting excess demand for money? does an increase in saving tend, holding income constant, instantaneously to depress the interest rate? In the flow equilibrium sense, the answer to both questions is yes; in the stock equilibrium sense, it is no.

To illustrate these points, I will review briefly a small subsequence of the literature on "loanable funds" and "liquidity preference" theories of the rate of interest. I have chosen only a few of the many articles to comment on, and I have tried to choose some which throw the issues raised in this paper into sharp relief.

The controversy began with an exchange between Ohlin and Keynes in the *Economic Journal* in 1937. There is no doubt that this exchange represents a mutual misunderstanding at various levels. Keynes continually argued that the interest rate was determined by the demand to hold money, while Ohlin referred to the interest rate as being determined in the market for flows of credit. Subsequent discussion fixed on the money versus credit dichotomy rather than on the stock versus flow dichotomy. As we have seen above, in both beginning-of-period and end-of-period equilibrium there is a symmetry between all assets markets. It is pointless to argue which asset market condition determines any particular price or rate of return: they determine all prices and rates of return simultaneously. The question of whether the interest rate is determined "in" the money market or the bond market is a red herring (at least for equilibrium analysis) in both versions.

However, I think it is clear that Ohlin had an "end-of-period" and Keynes a "beginning-of-period" vision of asset market equilibrium. Ohlin writes (1937, pp. 224–25):

> The willingness of certain individuals to *increase* their holdings of various claims and other kinds of assets *minus* the willingness of others to *reduce* their corresponding holdings gives the supply curves for the different kinds of new credit during the period. Naturally, the quantities each individual is willing to supply depend on the interest rate. . . . Similarly the total supply of *new* claims minus the reduction in the outstanding volume of *old* ones gives the demand—also a function of the rates of interest—for the different kinds of credit during the period. The prices fixed

> on the market for these different claims—and thereby the rates of interest—are governed by supply and demand in the usual way. [Italics in original]

This passage seems to me to be a good description of what I call "end-of-period" equilibrium. Keynes expressed himself more cryptically and never explicitly referred to the balance sheet constraint as a way of reconciling the money versus credit debate. His language, however, fits very well the model I have called stock equilibrium (Keynes 1937, p. 250):[6] "The resulting theory [i.e., Keynes's own], whether right or wrong, is exceedingly simple—namely that the rate of interest . . . has to be established at the level which, in the opinion of those who have the opportunity of choice—i.e., of wealth-holders—equalizes the attractions of holding idle cash and of holding the loan."

In succeeding years two schools seem to have grown up without completely recognizing the differences between them. Tobin (1969), for example, invariably and instinctively formulates the determination of asset prices as an equilibrium between stock demands and stock supplies of assets. Patinkin, on the other hand, devoted *Money, Interest, and Prices* to an elaboration of the end-of-period equilibrium model. That there is a certain lack of awareness that end-of-period and beginning-of-period equilibrium are distinct and inconsistent except under the assumption of perfect foresight is clear from passages like the following (Patinkin 1958, p. 302): "True the demand for money as a flow is quite different from the demand for money as a stock; in fact, they do not even have the same dimensions! Nevertheless, the *excess* demand for money as a flow is identical with the *excess* demand for money as a stock."

This statement about excess demands is true only under very restrictive conditions, as reference to equation (5a) will make clear. First, the deficit is clearly assumed to be zero by Patinkin: his economy is in a steady state. Second, there is no saving and no anticipated changes in prices so that $M^D(\Delta t) = M^D(0)$. Third, $k\Delta t$ is assumed to be 1, so that the only term left is the stock excess demand. The view that asset markets operate only to adjust stock disequilibrium appears to underlie this passage. Patinkin later writes (1958, p. 306): "Stock analysis, as well as flow analysis, presupposes a period of time: namely the period between the moment *at* which the individual is making his plans, and the moment *for* which he is making them. Hence if the periods presupposed by the analyses are the same, the excess-demand function of stock analysis must be identical with that of flow analysis. This proposition holds also in the limiting case where the period is instantaneous."

This passage seems to ignore the problem of choosing whether agents plan for the beginning or the end of the period in trading assets. It prejudges the issue by hiding in the words "if the periods presupposed by the analyses are the same" the idea that the analyses also identify the same part of the period as the moment for which the agent is making plans. Finally, Patinkin sums up his views (1958, p. 307): "It should also be clear from the above that the distinction that is sometimes attempted between flow equilibrium and stock equilibrium is an inadvisable one. Once again, if the periods are the same, the existence of one type of equilibrium implies the existence of the other." This way of talking ignores the distinction I have been trying to establish above.

---

[6] Keynes tried to patch things up between himself and Ohlin by introducing the planned flow of investment as a new argument (the finance motive) in the stock demand for money. This invention leaves the stock versus flow problem unresolved, of course.

A final passage from Johnson may be in order, since it seems to represent a widely accepted "compromise" view (Johnson 1961, pp. 6–7):[7]

> The liquidity preference–loanable funds debate turns on the question of whether the rate of interest is better regarded as equilibrating the flow of funds onto and off the market for securities or as equilibrating the demand for and supply of the stock of cash. The answer, which is now so deeply embedded in mathematical argument that no one can be sure he has got it right, seems to be that the stock-flow distinction is irrelevant, since either theory can be expressed in stock or flow terms; and that it makes no difference whether one works with money or securities, provided first, that one is concerned only with the determination of the equilibrium rate of interest, and, second, that one realizes that this is a general equilibrium problem which can be reduced only by artifice to a problem of equilibrium in one market. In more formal terms, if one assumes to begin with that the markets for goods and factors are in equilibrium, equality between the demand and supply for money implies equality between the (stock and flow) demand for and supply of loans and vice versa. The two theories become different, however, when applied to dynamic analysis of disequilibrium situations, since liquidity preference theory implies that the rate of interest rises only in response to an excess of the demand for over the supply of money, whereas loanable funds theory implies that it rises only in response to an excess supply of over demand for securities, and when the goods and factors markets are out of equilibrium an excess demand for money does not necessarily imply an excess supply of securities. In a dynamic context, the loanable funds theory definitely makes more economic sense; and the sustained resistance of Keynesians

---

[7] The last sentence is quoted (approvingly, given the context in which it appears) by Leijonhufvud (1968, p. 65, n. 13).

> to admitting it, evident most notably in the prolonged defense in the English literature of the proposition that an increase in the propensity to save lowers the interest rate only by reducing the level of income, is a credit to their ingenuity rather than their scientific spirit.

It may be worthwhile to point out in detail which arguments in this passage seem to be correct, and which do not, in light of the analysis of this paper. The stock flow distinction is relevant, even though in either stock or flow equilibrium either the money or securities market can be eliminated. (The remaining expressions will not be identical in the two versions, however.) Even if markets for goods and factors are not in equilibrium, equality between the stock demand for and stock supply of money implies equality between the stock demand for and supply of securities, and vice versa. Only if the goods and factor markets are in equilibrium does flow equilibrium *in the money market imply* flow equilibrium in the securities market. The two theories (stock and flow equilibrium) are different even in static analysis of equilibrium situations, as long as perfect foresight is not assumed.

In the last part of this passage, Johnson implicitly assumes the flow or end-of-period equilibrium model, and his remarks are correct in that context. If the economists he mentions in the last quoted sentence had beginning-of-period equilibrium in mind, perhaps they did not deserve so harsh a judgment on their scientific integrity.

Perhaps the only constructive way out of this confusion, short of developing more exact models of aggregate transactions, is for theorists to become familiar with both specifications, bring the choice between them up to the conscious level, and make explicit which vision underlies their result.

8. Conclusions

We can draw the following main conclusions from these arguments:

1. There exist two different ways of specifying asset equilibrium conditions in a macro model, each internally consistent both in period models and in continuous time.
2. The two specifications are consistent with each other only when there is perfect foresight on the part of agents concerning near future prices and rates of return.
3. There is no easy choice between the two versions of asset equilibrium, because they are both highly abstract approximations to a reality in which transactions are discrete and unsynchronized.
4. Certain theoretical propositions hold in one model but not in the other. For instance, "Walras's law" holds in end-of-period or flow

equilibrium between goods and asset markets. It does not apply in beginning-of-period or stock equilibrium, although in stock equilibrium it is possible to eliminate any asset market. For another instance, it is possible to have asset market equilibrium in the stock version without goods market equilibrium, so that an *IS-LM*-type division of the equilibrium conditions is natural. No such division is possible in flow equilibrium models.

These results lead me to think that the most fruitful line of future inquiry may be into models where the explicit structure of the transactions process is studied more precisely, rather than into "disequilibrium" variants of either of these models.

References

Foley, D., and Sidrauski, M. *Monetary and Fiscal Policy in a Growing Economy.* New York: Macmillan, 1971.

Hahn, F. "The Rate of Interest and General Equilibrium Analysis." *Econ. J.* 65 (March 1955): 52–66.

Johnson, H. G. "The General Theory after Twenty-five Years." *A.E.R.* 50 (May 1961): 1 ff.

Keynes, J. M. "Alternative Theories of the Rate of Interest." *Econ. J.* 47 (June 1937): 241–52.

Leijonhufvud, A. *On Keynesian Economics and the Economics of Keynes.* London: Oxford Univ. Press, 1968.

May, J. "Period Analysis and Continuous Analysis in Patinkin's Macroeconomic Model." *J. Econ. Theory* 2, no. 1 (March 1970): 1–9.

Ohlin, B. "Some Notes on the Stockholm Theory of Savings and Investment. II." *Econ. J.* 47 (June 1937): 221–40.

Patinkin, D. "Liquidity Preference and Loanable Funds: Stock and Flow Analysis." *Economica*, n.s. 25 (November 1958): 300–318.

Rybczynski, T. M. "Factor Endowments and Relative Commodity Prices." *Economica* 22 (November 1955): 336–41.

Tobin, J. "A General Equilibrium Approach to Monetary Theory." *J. Money, Credit and Banking* 1, no. 1 (February 1969): 15–29.

Covered Interest Arbitrage: Unexploited Profits?

Jacob A. Frenkel

University of Chicago and Tel-Aviv University

Richard M. Levich

New York University

Empirical studies of covered interest arbitrage suggest that the parity condition is not always satisfied and thus implying unexploited profit opportunities. This paper provides a procedure for estimating transaction costs in the markets for foreign exchange and for securities. Allowance for these costs accounts for most of the apparent profit opportunities. It is shown that in addition to transaction costs, demand and supply elasticities in the various markets and lags in executing arbitrage can account for all of the apparent profit opportunities. It is concluded that empirical data are consistent with the interest parity theory and that covered interest arbitrage does not entail unexploited profit opportunities.

The interest parity theory maintains that in equilibrium the premium (or discount) on a forward contract for foreign exchange is related to the interest rate differential according to

$$\frac{F - S}{S} = \frac{i - i^*}{1 + i^*}, \tag{1}$$

where F and S are, respectively, the forward and spot exchange rates, i is the domestic rate of interest on a particular class of securities, and i^* is the foreign rate of interest. The rates of interest should satisfy the

---

We are indebted to Dan Galai, Jo Anna Gray, Roy J. Ruffin, and an anonymous referee for helpful comments on an earlier draft.

[*Journal of Political Economy*, 1975, vol. 83, no. 2]

© 1975 by The University of Chicago. All rights reserved.

comparability criterion: the pair of securities should be identical in all respects (e.g., maturity, risk class, etc.) except for the currency of denomination. When equation (1) is satisfied there are no profit opportunities from covered interest arbitrage.

The notion that arbitrage activity will eliminate unexploited profit opportunities in the market for foreign exchange is, of course, not new and can be found in Ricardo's *Reply to Mr. Bosanquet:*

> [T]heory takes for granted, that whenever enormous profits can be made in any particular trade, a sufficient number of capitalists will be induced to engage in it, who will, by their competition, reduce the profits to the general rate of mercantile gains. It assumes that in the trade of exchange does this principle more especially operate; it not being confined to English merchants alone, but being perfectly understood, and profitably followed, by the exchange and bullion merchants of Holland, France and Hamburgh; and competition in this trade being well known to be carried to its greatest height. [Ricardo 1811, pp. 9–10]

Empirically, however, the parity condition is not always satisfied. Recent studies have analyzed the persisting deviations from the interest parity line in terms of transaction costs (Branson 1969), political risk (Aliber 1973), and capital market imperfections (Prachowny 1970; Frenkel 1973). A wider survey of the issues was provided by Stein (1962) and Officer and Willet (1970).

The purpose of this paper is to combine the various explanations into a comprehensive framework and to provide the empirical counterpart to those explanations. It is shown that observed data are consistent with the hypothesis that the relevant markets are highly efficient in eliminating unexploited profit opportunities. The nature of the transaction costs is explored in Section I. An empirical estimate of these costs is provided in Section II. These estimates are then used in Section III to account for some of the deviations from the interest parity line; in this context we also investigate the comparability criterion. The role of demand and supply elasticities in the securities and foreign exchange markets and the role of timing the execution of arbitrage are analyzed in Section IV, with concluding remarks in Section V.

I. Transaction Costs and the Interest Parity Theory

As is well known, transaction costs in securities and foreign exchange markets imply the existence of a neutral band—around the interest parity line—within which no additional arbitrage is profitable. Thus, points which are off the traditional interest parity line may still be interpreted as "equilibrium" points in the sense of precluding arbitrage incentives, since within this band transaction costs exceed arbitrage profits.

There are basically two elements to the cost of transactions: (i) costs of transacting in the foreign exchange markets (spot and forward) and (ii) costs of transacting in the securities markets (domestic and foreign).

Consider first an outflow of covered arbitrage funds from the domestic to the foreign market. There are four transactions and thus four distinct transaction costs in the covered outflow: (*a*) sale of domestic securities with transaction costs of t percent, (*b*) purchase of foreign currency (spot) with transaction costs of t_S percent, (*c*) purchase of foreign securities with transaction costs of t^* percent and, (*d*) sale of foreign currency (forward) with transaction costs of t_F percent. The cost (C) associated with a capital outflow of the amount χ is the foregone earnings on the holdings of domestic securities:

$$C = \chi(1 + i). \tag{2}$$

The revenue (R) derived from a covered investment of these funds in comparable securities abroad is

$$R = \chi\Omega(1 + i^*)F/S, \tag{3}$$

where

$$\Omega = (1 - t)(1 - t_S)(1 - t^*)(1 - t_F).$$

Equating the marginal cost (from [2]) with the marginal revenue (from [3]) yields the *lower* limit on $p[p \equiv (F - S)/S]$ for which marginal *outflow* of arbitrage funds is profitable:

$$\underline{p} = \frac{(1 + i) - \Omega(1 + i^*)}{\Omega(1 + i^*)}. \tag{4}$$

This formulation of transaction costs assumes that there are no lumpy costs independent of the quantity transacted and that the initial position of arbitragers is in securities rather than in cash. Since, however, at each point in time some fraction of the existing stock of securities matures, some arbitragers may initially hold cash. In this case their transactions costs are lower.[1] Consequently, to have a more conservative estimate of the cost of transactions we replace Ω in equation (4) and equation (7) below by Ω' where $\Omega' = \Omega/(1 - t)^2$.

Consider next an inflow of covered arbitrage funds from the foreign to the domestic market. Again, four transactions are involved in the covered inflow: (*a*) sale of foreign securities with transaction costs of t^* percent, (*b*) sale of foreign currency (spot) with transaction costs of t_S

---

[1] In this case arbitragers equate the marginal return from buying domestic securities $(1 - t)(1 + i)$ with the marginal return from a covered purchase of foreign securities $(1 - t_S)(1 - t^*)(1 - t_F)(1 + i^*)F/S = [\Omega/(1 - t)](1 + i^*)F/S$. We are indebted to Rudiger Dornbusch for raising this issue.

percent, (*c*) purchase of domestic securities with transaction costs of t percent, and (*d*) purchase of foreign currency (forward) with transaction costs of t_F percent. Therefore, the cost (foregone earnings) and revenue of the operation are, respectively,

$$C = \chi(1 + i^*), \tag{5}$$

$$R = \chi\Omega(1 + i)S/F. \tag{6}$$

Equating marginal cost with marginal revenue (from [5] and [6], respectively) yields the *upper* limit on p for which marginal *inflow* of arbitrage funds is profitable:

$$\bar{p} = \frac{\Omega(1 + i) - (1 + i^*)}{1 + i^*}. \tag{7}$$

Conditions (4) and (7) define a neutral band in which marginal covered arbitrage flows are not profitable. The values of p bounded in this band are

$$\bar{p} \leq p \leq \underline{p}. \tag{8}$$

For the special case in which all transaction costs are zero, condition (8) reduces to equation (1) (as $\bar{p} = p = \underline{p}$), and the neutral band includes only the traditional interest parity line. Whenever any of the costs of transaction are positive, $\bar{p} < \underline{p}$.

II. Estimates of the Cost of Transactions

Since there is no direct measure of the cost of transactions in the foreign exchange and securities markets, they must be estimated indirectly. One method to estimate the cost of transactions in the markets for foreign exchange is to study the behavior of triangular arbitrage.

The essence of triangular arbitrage is to keep cross-exchange rates consistent. Accordingly, in the absence of transaction costs, consistency of equilibrium requires that

$$(\$/£)_\tau = (\$/\mathrm{DM})_\tau(\mathrm{DM}/£)_\tau, \tag{9}$$

where the left-hand side of (9) indicates the price of 1 pound sterling in terms of U.S. dollars, and the right-hand side indicates the product of the price of one German mark in terms of U.S. dollars and the price of 1 pound sterling in terms of German marks. The subscript τ indicates that those prices are for foreign exchange delivered at the same maturity τ (Grubel 1966). If, however, transaction costs are not absent, then the upper limit of the absolute discrepancy between the two sides of equation (9) should correspond to the cost of transacting. Accordingly, this is the way we estimate the cost of transacting in foreign exchange. This estimate includes brokerage fees, the cost of being informed, and all other costs associated with the foreign exchange transaction.

The empirical analysis covers the period between January 1962 and November 1967. The choice of the period was determined by our preference for a relatively "quiet period" which was terminated by the British devaluation. Using weekly observations of the spot and of the (90-day) forward exchange rates, we have computed two series of 307 weekly percentage deviations from the triangular arbitrage of equation (9) for the spot and for the forward exchange rates.[2] The upper limits of these series of deviations correspond to t_S and t_F, respectively.[3] However, to allow for errors in measurements and other noise in data we have taken a more conservative measure: our estimates of t_S and t_F are the percentage deviation which bounds 95 percent of the deviations from triangular arbitrage.

The arbitrage between pound sterling and U.S. dollar in equation (9) was assumed to take place through the German mark, although some other currency might be used as the intermediate vehicle. It is presumed, however, that competition will assure that at the margin the cost of transactions will tend to be equalized among the various major currencies. To examine this presumption, we have computed these costs for the case in which the Canadian dollar ($C) replaces the German mark as the intermediate vehicle according to equation (9′). In addition, we have computed the cost of arbitrage between a different pair of currencies—the U.S. dollar and the Canadian dollar—for the case in which the pound sterling serves as the intermediate currency as in equation (9″):

$$(\$/£)_\tau = (\$/\$C)_\tau(\$C/£)_\tau \qquad (9')$$

$$(\$/\$C)_\tau = (\$/£)_\tau(£/\$C)_\tau \qquad (9'')$$

Table 1 summarizes the various estimates of the transaction costs in the spot and the 90-day forward markets for foreign exchange; these estimates correspond to the percentage cost which bounds 95 percent of the deviations from the triangular arbitrage of equations (9), (9′), and (9″). As can be seen, the estimates are almost invariant with respect to the choice of the intermediate currency as well as with respect to the choice of the pair of currencies between which arbitrage takes place.

---

[2] Exchange rates involving the U.S. dollar are closing bid price for the day (International Monetary Fund. This source published daily data for the period under study). The DM/£ and Canadian $/£ rates are midpoints of the daily range (Samuel Montagu & Co.).

[3] In a strict sense, the upper limit of these deviations for the whole period would correspond to the cost of transactions only if the structure of these costs remained stable. Although there is no formal statistical test for stability, we have examined these deviations for arbitrarily chosen subperiods and concluded that the structure of the deviations seems to be stable over the period as a whole.

TABLE 1

ESTIMATES OF THE PERCENTAGE COST OF TRANSACTIONS IN THE MARKETS FOR FOREIGN EXCHANGE (SPOT AND 90-DAY FORWARD)

Arbitrage between	Intermediate Currency	t_S	t_F	$t_S + t_F$
\$U.S. and £.....	DM	0.051	0.076	0.127
\$U.S. and £.....	\$C	0.058	0.068	0.126
\$U.S. and \$C....	£	0.058	0.073	0.131

To estimate the cost of transactions in the market for securities (t), we follow Demsetz (1968) in arguing that these costs are comprised mainly of two elements: the ask-bid spread and the brokerage fee. The ask-bid spread measures the reward that specialists must obtain for standing ready to buy and sell on *immediate* demand; it measures the price of immediacy in both buying and selling securities, and thus under competitive conditions the ask-bid spread corresponds to the cost of an immediate round-trip exchange (Demsetz 1968). In estimating the total cost of transacting in securities we adopt Demsetz's estimate according to which total costs are about 2.5 times the ask-bid spread. Accordingly, we have computed the ask-bid spread for the monthly observations of the 90-day U.S. Treasury bill.[4] The modal value of the spread is 0.0076 percent, and thus the modal value of total transaction costs in the U.S. securities market (including brokerage fee) is 0.019 percent.[5] In the absence of information on the ask-bid spreads on the U.K. and the Canadian treasury bills, we have assumed that those are equal to the spreads on the U.S. Treasury bill and thus that $t = t^*$.[6] Adding the cost of transactions in securities to the cost of transacting in foreign exchange yields the total cost of transactions in a covered interest arbitrage.[7] Table 2 presents estimates

---

[4] The bid price and the ask price (expressed as a percentage of par) were computed according to (i) bid price = 100 − (bid yield × days to maturity)/360, (ii) ask price = 100 − (ask yield × days to maturity)/360 and the percentage spread is (ask price − bid price)/ask price. The data on the bid yield and ask yield are from *Solomon Brothers Monthly Bond Report*. It should be noted that, since 90-day Treasury bills are not issued on every day, for each observation date we have computed the ask-bid spread on the Treasury bill whose maturity was closest to 90 days. The range of maturities accepted was between 87 and 93 days.

[5] The distribution of the spreads around the modal values of the spread is extremely tight; in fact, in 52 of the 71 months under investigation, the spreads were between 0.0072 and 0.0078 percent.

[6] Since the modal value of t is much smaller than the other components of the cost of covered interest arbitrage, it seems that this approximation does not affect the results significantly.

[7] In principle the ask-bid spread could also be used to estimate transaction costs in the foreign exchange markets. A small sample of data on the ask-bid spread in foreign exchange transactions yields estimates that are of the same orders of magnitude as our estimates. Likewise, for the period under study, the IMF published data on the ask-bid spread for the spot pound sterling; these data also yield similar estimates to ours. However lack of data on the various currencies for the whole period prevented a complete comparison of the two alternative methods.

TABLE 2

TOTAL TRANSACTION COSTS IN 90-DAY COVERED INTEREST ARBITRAGE

$t_\$$ and t_F Estimated by	Cost of Transactions (%)
DM	0.146
$C	0.145
£	0.150

of the total cost using the various estimates of table 1 together with the modal value of the transaction costs in securities.[8]

The estimates in table 2 are rather close to Branson's (1969) estimate of 0.18 percent. They are bounded by Holmes and Schott's (1965) estimate of 0.25 percent and Einzig's (1961) estimates of 0.03–0.10 percent; however, they are much lower than Keynes's (1923) estimate of 0.50 percent.

III. Transaction Costs, Comparability and Covered Interest Arbitrage

The model that was developed in Section I is a one-period model where the rates of interest and the forward premiums are those that are relevant for the specified period. Published sources of data report their *annualized* values. Since the rates of interest are computed according to the bankers' discount formula, the process of annualizing the 90-day yields and the 90-day forward premiums amounts to quadrupling the 90-day magnitudes.[9] Since there is no presumption that the forward premium on a yearly contract for foreign exchange should be four times the forward premium on a 90-day contract, it is clear that annualizing the premiums might introduce noise to the data. Similar noise might be introduced by the annualized interest rates series since there is no presumption that the interest differential on 1-year securities should be four times the corresponding differential on 90-day securities.[10]

As a purely arithmetical matter, it is clear that the method of annualization multiplies all deviations from the interest parity line by about four and thereby creates illusory profit margins. Moreover, since our estimates of the cost of transactions are for 90-day covered arbitrage, applying these estimates to the annualized data raises the question of the dependence of the cost of transactions on the maturity of the arbitraged assets.[11] Consequently, to pay adherence to the 90-day, one-period model, we have converted the annualized data back into their 90-day counterparts by dividing the published premium and interest rates by four.[12] The following analysis is based on the transformed data.

---

[8] When the initial position of arbitragers is in securities, the estimates in table 2 are conservative since it assumes that arbitragers who are the most efficient in the currency markets are also the most efficient in the securities market.

[9] The bankers discount formula is $i = (100 - P)360/D$ where i is the bankers discount yield, P the bond price as a percentage of par, and D the number of days to maturity. For a discussion on the data see *Federal Reserve Bulletin* (October 1964).

[10] For an analysis of the maturity structure of the forward exchange rate and the consequences of different yield curves see Grubel (1966, chaps. 8, 13). We are indebted to Herbert G. Grubel for a discussion on this point.

In incorporating the costs of transacting into the analysis of covered interest arbitrage, a basic question is which pair of securities should be used. Traditionally, when arbitrage flows between the United States and the United Kingdom were considered, the securities used were the 90-day U.S. and U.K. treasury bills. This choice raises the question of whether this pair is comparable in terms of risk of default. Recently, Aliber (1973) has argued that the traditional pair does not satisfy the comparability criterion since these securities are issued in financial centers that differ in political risk. Instead, he argues, one should go to an external market and compare the rates of interest on a (U.S.) dollar-denominated security issued in (say) Paris with a (U.K.) pound-denominated security that is also issued in Paris. Since these securities are issued in an external center, they are more comparable.

To examine this proposition we have used two pairs of securities (i) the traditional pair: weekly U.S. and U.K. 90-day treasury bills[13] and (ii) the external pair: weekly Euro-market rates on 90-day securities denominated in U.S. dollars and U.K. pounds.[14]

---

[11] Due to the lack of data for the period in question we could not establish the exact relationship between the estimates of the cost of transactions and the length of the contract. However, for the latter period (1972–74) an examination of the ask-bid spread in foreign exchange transactions reveals that the spread for a 12-month forward contract exceeds the spread for the 3-month contract. For the relevant data on the latter period see *International Monetary Market Year Book, 1972–1973* and the *Daily Information Bulletin* published by the staff of the International Monetary Market.

[12] As a theoretical matter, one would expect arbitragers to compare the forward premium to the (discounted) interest rate differential where the interest rates are computed as a holding period yield. Since the bankers discount yield is systematically below the holding period yield, use of the former in traditional calculations of arbitrage incentives introduces a bias. As an empirical matter, however, the difference between the two methods of interest rate calculation becomes significant only for high interest rates. Calculations using holding period yield do not effect significantly the results in table 3 below.

[13] All 3-month Treasury bill rates are taken from the *Federal Reserve Bulletin*, various issues. In fact, since the *Federal Reserve Bulletin* reports the latest 3-month rate, some of the rates used are for bills that are slightly less than 3-months by 2 or 3 days. As was explained in the text, the annualized data were transformed back into their original 90-day values.

[14] The external rates used are Euro-sterling deposits in Paris and Euro-dollar deposits in London; both are 3-month middle closing rates (*Bank of England Quarterly Bulletin*, various issues). Ideally, one would prefer using dollar and sterling deposits issued in the same external center. However, due to lack of data we follow Aliber's procedure (1973) by using sterling deposits in Paris and dollar deposits in London. Aliber has argued that this compromise does not alter the results to a significant extent.

TABLE 3

TRANSACTION COSTS AND NEUTRAL BAND FOR 90-DAY COVERED ARBITRAGE

		OBSERVATIONS BOUNDED WITHIN NEUTRAL BAND (%)			
		Traditional Pair		External Pair	
ARBITRAGE BETWEEN SECURITIES DENOMINATED IN	t_S AND t_F ESTIMATED AS	Ω	Ω'	Ω	Ω'
$U.S. and £.......	DM	87.0	82.4	99.7	99.3
$U.S. and £.......	$C	87.0	82.1	99.7	99.3
$U.S. and $C......	£	88.3	83.4	N.A.	N.A.

An additional proposition is that, other things equal, the cost of transactions should account for a similar fraction of the deviations from the interest parity line independently of the currency denomination of the arbitraged securities. This proposition is tested by comparing the role of transaction costs in accounting for deviations from the parity line of a covered interest arbitrage between U.S. and U.K. treasury bills with their role in accounting for the deviations when the arbitrage is between U.S. and Canadian treasury bills. Due to lack of data on external Canadian interest rates, a similar comparison could not be applied to the external pair.

Applying the formula developed in Section I and using the estimates of the transaction costs in the foreign exchange market (t_S, t_F) and the modal estimate of the transaction costs in the securities market (t), we have computed the neutral band around the interest parity line.

Table 3 summarizes the results of these computations for the traditional and the external pairs of securities using the alternative estimates of the cost of transactions from table 2, assuming two alternative initial positions corresponding to Ω and Ω'.

As is seen from table 3, when the traditional pair of securities is used the cost of transactions accounts for about 85 percent of the apparent profit opportunities implied by deviations from the interest parity line. It is noteworthy that the proportion of the deviations from the U.S.-U.K. parity line explained by the cost of transactions is almost identical with the proportion of the deviations from the U.S.-Canada parity line explained by these costs. The role of the cost of transactions is even more striking when applied to the external pair of securities, in which case they account for almost all of the deviations from the interest parity line.

Given that so many of the apparent profit opportunities can be explained by transactions cost, very little is left to be explained. In the following section, however, we elaborate on two additional explanations.

IV. Accounting for the Residual

A. The Role of Elasticities

The formulation of the upper and lower bounds which generate the neutral band around the interest parity line presumes that the rates of interest are independent of the amounts that are borrowed and lent and that the price of foreign exchange (spot and forward) is independent of the quantity of foreign exchange which is bought and sold. Once these assumptions are relaxed the marginal cost and revenue of covered interest arbitrage would depend on the elasticities of the demand and the supply in the securities and foreign exchange markets. The existence of elasticities which are less than infinite will widen the neutral band and will therefore account for a larger percentage of the deviations from the line. The lower limit on p for which marginal outflow of arbitrage funds is profitable becomes

$$\underline{p} = \frac{(1 + \lambda i) - \Omega[1 + \sigma^* i^* - \Psi(1 + i^*)]}{\Omega[1 + \sigma^* i^* - \Psi(1 + i^*)]}, \tag{4'}$$

where

$$\lambda = 1 + 1/\varepsilon, \sigma^* = 1 - 1/\eta^*, \Psi = 1/\eta_F + 1/\varepsilon_S$$

and where

$\varepsilon \equiv (\partial \chi_b / \partial i)(i/\chi_b) > 0$ is the elasticity of the domestic supply of funds (where χ_b denotes quantity borrowed),

$\eta^* \equiv -(\partial \chi_\ell / \partial i^*)(i^*/\chi_\ell) > 0$ is the elasticity of the foreign demand for funds (where χ_ℓ denotes quantity lent),

$\eta_F \equiv -(\partial \chi / \partial F)(F/\chi) > 0$ is the elasticity of the domestic demand for (forward) foreign exchange, and

$\varepsilon_S \equiv (\partial \chi / \partial S)(S/\chi) > 0$ is the elasticity of the domestic supply of (spot) foreign exchange.

The upper limit on p for which marginal inflow of arbitrage funds is profitable becomes

$$\bar{p} = \frac{\Omega[1 + \sigma i - \zeta^*(1 + i)] - (1 + \lambda^* i^*)}{1 + \lambda^* i^*}, \tag{7'}$$

where

$$\lambda^* = 1 + 1/\varepsilon^*, \sigma = 1 - 1/\eta, \zeta^* = 1/\eta_S^* + 1/\varepsilon_F^*,$$

and where

$\varepsilon^* \equiv (\partial \chi_b / \partial i^*)(i^*/\chi_b) > 0$ is the elasticity of the foreign supply of funds,

$\eta \equiv -(\partial\chi_d/\partial i)(i/\chi_d) > 0$ is the elasticity of the domestic demand funds,

$\eta_S^* \equiv -(\partial\chi/\partial S^*)(S^*/\chi) > 0$ is the elasticity of the foreign demand for (spot) foreign exchange, and

$\varepsilon_F^* \equiv (\partial\chi/\partial F^*)(F^*/\chi) > 0$ is the elasticity of the foreign supply of (forward) foreign exchange.

As is clear, for the special case in which all the elasticities are infinite, equations (4′) and (7′) become identical with equations (4) and (7), and when, in addition, all the cost of transactions are zero, equation (1) is obtained.

Given that a significant portion of the deviations from the interest parity line are explained by the cost of transacting, we have computed (using equations [4′] and [7′]) the maximal values of the elasticities which together with the estimated cost of transactions generate a neutral band within which all of the observations are bounded and thus all of the apparent profit opportunities are accounted for. In the absence of prior knowledge we have assumed that all of the elasticities are equal. These computations revealed that, given the cost of transactions, the values of the elasticities needed to account for all of the deviations from the interest parity line are very high, reaching many hundreds in all cases. These high values are not inconceivable given the highly competitive organization of the markets in question.[15]

B. *The Role of Timing*

A second possible explanation accounting for the residual profit opportunities concerns the timing of transactions. To the extent that there exists a time interval between the receipt of information signaling arbitrage profit opportunities and the actual execution of the arbitrage transactions, prices might change to bid away profits. Since prices may change, it follows that quoted prices may differ from actual transaction prices and thus that covered arbitrage is not risk free. The fact that ex post computations reveal the existence of profits on some dates does not imply that an

[15] In a previous paper (Frenkel 1973), a similar method was used to compute the elasticities of the demand and the supply of funds without allowing for transaction costs. The computed elasticities were extremely low. What the present analysis suggests is that the value of the implied elasticities is much higher once we incorporate the estimates of the cost of transactions and is therefore consistent with the notion of the high degree of competitiveness. It should be noted, however, that the association between the organization of the market and the elasticities of demand and supply faced by individual arbitragers is not immediate. The fact that individuals face an upward-sloping supply of funds might reflect risk factors and other characteristics of the quality of the loan rather than monopolistic markets. However, high aggregate elasticities are consistent with high degrees of competitiveness. We are indebted to the referee for raising this issue. (The editor has informed us that a similar issue was raised in an unpublished paper by E. Tsoa.)

TABLE 4

MEAN PERCENTAGE PROFITS AND LAGS IN TRANSACTIONS

ARBITRAGE BETWEEN SECURITIES DENOMINATED IN	t_s AND t_F ESTIMATED AS	N		t Mean Profit		$t+1$ Mean Profit	
		Ω	Ω'	Ω	Ω'	Ω	Ω'
$U.S. and £	DM	40	54	0.055 (7.2)	0.058 (8.4)	0.038 (3.7)	0.045 (5.4)
$U.S. and £	$C	40	55	0.056 (7.3)	0.058 (8.4)	0.039 (3.8)	0.044 (5.4)
$U.S. and $C	£	36	51	0.056 (7.4)	0.057 (8.4)	0.044 (3.7)	0.045 (4.8)

NOTE.—t-statistic in parentheses.

investor could have designed an ex ante trading rule which yields sure profits.

To examine this possibility, we have studied the consequences of a simple trading rule according to which investors calculate a profit signal at period t and execute their transactions at period $t + 1$. Table 4 summarizes the results of applying this trading rule to the traditional pair of securities. If investors were able to transact in quoted prices at period t, there would be N profit opportunities with a mean rate of return of about 0.05 percent (with a t-statistic in parentheses below the mean). However, when transactions are executed at period $t + 1$, some of the unexploited profit opportunities prove to be illusory, and the mean rate of profits is reduced to about 0.04 percent.[16] While as many as 30 percent of the returns in period $t + 1$ may be negative, the mean rate of return for the entire sample is still positive. It is especially noteworthy that, in addition to a smaller mean return, the returns in period $t + 1$ exhibit larger variance than those of period t and thus indicate greater risk. Ideally, a test of this trading rule should be based on data for consecutive transactions; since the results in table 4 are based on weekly data, they should be viewed only as indicative.[17]

[16] In addition to the role of timing, quoted prices might differ from transaction prices due to the method used in computing the former. The data used in the computations were taken from the *Federal Reserve Bulletin*. The Federal Reserve uses a sampling and averaging technique in reporting forward premium and interest rates. Therefore, the data base is what Holmes and Schott (1965) refer to as "consensus" data. It cannot be inferred that investors actually transacted and therefore made profits at these reported prices. The misspecification introduced by using the midpoint quotation (between bid and offer prices) is analyzed in detail by Agmon and Bronfeld (1975), who argue that errors in measurements and inaccurate definitions account for many of the apparent profit opportunities.

[17] Galai (1974) who suggested this approach has tested a similar trading rule for arbitrage between the Chicago Board Option Exchange and the New York Stock Exchange. Using daily and hourly data, he has shown that following this rule results in profits which do not differ significantly from zero.

V. Concluding Remarks

The essence of the interest parity theory is in establishing a relationship between the premium (or discount) on a forward contract for foreign exchange and the differential in interest rates on securities that are identical in all respects except for the currency of denomination. In principle, points which are off the traditional interest parity line can be interpreted as being equilibrium points if there exists a neutral band around the traditional line within which no covered arbitrage is profitable. Such a band is implied if there are transaction costs and/or if some (or all) of the elasticities of the demand and the supply (of securities and foreign exchange) are less than infinite. In addition, empirical deviations from the interest parity line can be the result of analyzing the wrong pair of securities and thus violating the comparability criterion.

The purpose of this paper was to explore the nature of the various transaction costs and to provide a novel procedure for estimating their magnitude. The estimated costs account for a very similar proportion of the deviations from the interest parity line when applied to different pairs of securities—the U.S.-Canada and the U.S.-U.K. pairs of treasury bills. In addition, to explore the role of the comparability criterion as a possible explanation for the persisting deviations from the parity line we have examined two pairs of securities—the traditional pair and an external pair. It was shown that allowance for the cost of transactions accounts for most of the apparent deviations even when we use the traditional pair of securities. The residual number of observations that are not accounted for could be due to many reasons: differential tax treatment, differential risk, governmental controls, elasticities of demand and supply that are less than infinite, time differential between observing a profit opportunity and executing the arbitrage activity, etc. We have examined the last two of the above reasons as possible sources for the residual deviations. The computed elasticities are extremely high and are consistent with the notion that these markets are highly competitive. Likewise, allowing for a lag between observing a profit opportunity and executing the transaction, reduces the number of apparent profit opportunities.

We conclude, therefore, by stating that the empirical data are consistent with the interest parity theory in the sense that covered interest arbitrage does not seem to entail unexploited opportunities for profit.

References

Agmon, Tamir, and Bronfeld, Saul. "The International Mobility of Short-Term Arbitrage Capital." *J. Bus. Finance and Accounting* (Spring 1975), forthcoming.

Aliber, Robert Z. "The Interest Rate Parity Theorem: A Reinterpretation." *J.P.E.* 81, no. 6 (November/December 1973): 1451–59.

Bank of England. *Bank of England Quarterly Bulletin*, various issues.

Branson, William H. "The Minimum Covered Interest Differential Needed for International Arbitrage Activity." *J.P.E.* 77, no. 6 (November/December 1969): 1028–35.

Demsetz, Harold. "The Cost of Transacting." *Q.J.E.* 82, no. 1 (February 1968): 33–53.

Einzig, Paul A. *A Dynamic Theory of Forward Exchange.* London: Macmillan, 1961.

Federal Reserve System. *Federal Reserve Bulletin,* various issues.

Frenkel, Jacob A. "Elasticities and the Interest Parity Theory." *J.P.E.* 81, no. 3 (May/June 1973): 741–47.

Galai, Dan. "Pricing of Options and the Efficiency of the CBOE Option Market." Mimeographed. Univ. Chicago, 1974.

Grubel, Herbert G. *Forward Exchange, Speculation, and the International Flow of Capital.* Stanford, Calif.: Stanford Univ. Press, 1966.

Holmes, A. R., and Schott, F. H. *The New York Foreign Exchange Market.* New York: Federal Reserve Bank, New York, 1965.

International Monetary Fund. *International Financial Statistics.* Washington: IMF. Monthly issues, 1962–67.

International Monetary Market. *International Monetary Market Year Book* 1972–1973. Chicago, 1973.

Keynes, John M. *A Tract on Monetary Reform.* London: Macmillan, 1923.

Montagu, Samuel & Co. *Montagu Monthly Review.* London, various issues.

Officer, Lawrence H., and Willet, Thomas D. "The Covered-Arbitrage Schedule: A Critical Survey of Recent Developments." *J. Money, Credit and Banking* 2, no. 2 (May 1970): 247–57.

Prachowny, Martin F. "A Note on Interest Parity and the Supply of Arbitrage Funds." *J.P.E.* 78, no. 3 (May/June 1970): 540–45.

Ricardo, David. *Reply to Mr. Bosanquet's Practical Observations on the Report of the Bullion Committee.* London, 1811.

Solomon Brothers. *Solomon Brothers Monthly Bond Report.* New York, various reports.

Stein, Jerome L. *The Nature and Efficiency of the Foreign Exchange Market.* Essays in International Finance, no. 40. Princeton, N.J.: Princeton Univ. Press, 1962.

Strategic Considerations in the Richardson Model of Arms Races

---

Michael D. Intriligator

University of California, Los Angeles

Parameters defining Richardson-type reaction functions for missile stocks are derived, for both deterrers and attackers, as explicit functions of certain underlying technical and timing factors in a simulated missile war. With two deterrers there generally exists a stable equilibrium of mutual assured destruction, where the equilibrium missile stocks depend on technical and timing factors; with two attackers an unstable equilibrium generally exists; and with a deterrer and an attacker no equilibrium exists. Above the stable equilibrium there exists a cone of mutual deterrence, while below it there exist regions of war initiation. Arms control and disarmament measures are treated using this framework.

1. Introduction

The Richardson model is one of the first analytic approaches to an arms race between two countries. In the Richardson model the dynamics of the arms races are summarized by a system of linear differential equations.[1] This model has been applied in modern treatments to arms races in the missile age.[2] One of the principal limitations of the Richardson model, however, is that it looks at the arms race from the outside as a mechanistic model rather than from the inside in terms of decisions made by defense planners.

The original version of this paper was presented at the Econometric Society meetings, New Orleans, 1971. The author is indebted to D. L. Brito and an anonymous referee for useful comments and suggestions. This research was completed under a grant from the National Science Foundation.

[1] See Richardson (1939, 1960) and Rapoport (1957, 1960).

[2] See Burns (1959), Boulding (1962), Kent (1963), McGuire (1965), Schelling (1966), Saaty (1968), and Pitman (1969). Most of these modern treatments apply the Richardson model to stocks of weapons, as done here, rather than to strategic budgets, as in the original Richardson development of the model.

[*Journal of Political Economy*, 1975, vol. 83, no. 2]

© 1975 by The University of Chicago. All rights reserved.

The purpose of this paper is to consider the arms race in terms of decisions made by defense planners. It derives the parameters of the reaction functions of the Richardson model as explicit functions of technical and timing factors. This derivation is based on a simulated war, represented by a dynamic model of a missile war, which can be utilized by strategic decision makers to analyze interactions in warfare. The basic premise of this paper is that any analytic study of arms must be based on strategic considerations regarding targets and rates of fire in the course of a *potential* war. Thus, the decision to acquire a weapon must be based on an evaluation of its effectiveness in the event of a war. Given certain assumptions as to the objectives of the two countries and the initiation of a potential war, the dynamic model implies a Richardson-type reaction of each country to the level of weapons held by its opponent. Of especial interest is the existence of equilibrium levels of weapons, which may be stable or unstable depending on the objectives of the defense planners in the two countries.

Section 2 summarizes the Richardson model, while section 3 summarizes the dynamic model of a missile war. Section 4 uses this dynamic model to treat simulated war outbreak. Sections 5 and 6 treat strategy and arms races in the case of two deterrers and in the case of attackers and asymmetric cases, respectively, presenting equilibrium levels of missile stocks, where they exist, as explicit functions of technical and timing factors in a simulated missile war. Section 7 presents applications to arms control and war imitation, and section 8 summarizes conclusions of the paper.

2. The Richardson Model

In the Richardson model two countries, A and B, are acquiring stocks of weapons, missiles, where $M_A(t)$ and $M_B(t)$ are missile stocks at time t in A and B, respectively. The Richardson model is the pair of linear differential equations

$$\dot{M}_A = a_1 M_B - a_2 M_A + a_3 \qquad (a_1, a_2 > 0), \tag{1}$$

$$\dot{M}_B = b_1 M_A - b_2 M_B + b_3 \qquad (b_1, b_2 > 0), \tag{2}$$

giving the time rate of change of missile stocks in each country as a linear function of the stock of missiles held by both countries. Each of the three terms on the right-hand sides of these equations has an interpretation. The first is the "defense term," indicating the positive effects of the opponent's missiles on the acquisition of missiles. The second term is the "fatigue" term, indicating the negative effects of one's own missiles on the acquisition of missiles due to the burden of holding missiles. The third term is the "grievance" term, indicating the effects of all other factors on the acquisition of missiles. The basic Richardson assumption is that these causes of arms races have independent and additive effects on the acquisition (or destruction) of missiles.

For an equilibrium there is no change in missile stocks, that is, both $\dot{M}_A$ and $\dot{M}_B$ are zero, yielding reaction functions[3]

$$M_A = a'_1 M_B + a'_3, \quad \text{where} \quad a'_1 = \frac{a_1}{a_2}, \quad a'_3 = \frac{a_3}{a_2}, \tag{3}$$

$$M_B = b'_1 M_A + b'_3, \quad \text{where} \quad b'_1 = \frac{b_1}{b_2}, \quad b'_3 = \frac{b_3}{b_2}. \tag{4}$$

These reaction functions give desired numbers of missiles for each country as functions of the number of missiles held by the opponent. The essence of the Richardson model is that this dependence is linear.[4]

Assuming the grievance terms a_3 and b_3 are positive, an equilibrium exists if the product of the slopes in (3) and (4) is less than unity,

$$a'_1 b'_1 < 1. \tag{5}$$

In this case the equilibrium is stable, that is, small movements away from the equilibrium would set in motion forces that, via the basic equations of the model, would restore the equilibrium. The equilibrium missile stocks are then

$$M_A^E = \frac{a'_1 b'_3 + a'_3}{1 - a'_1 b'_1}, \tag{6}$$

$$M_B^E = \frac{a'_3 b'_1 + b'_3}{1 - a'_1 b'_1}, \tag{7}$$

and the system tends toward this equilibrium.

If the product of the two slopes in (5) is greater than unity there could be an equilibrium provided the grievance terms were negative, but in this case the equilibrium would be unstable: there would be a downward spiral of missiles (toward zero levels) for slight movements below the equilibrium and an upward spiral of missiles for slight movements above the equilibrium.[5]

Sections 5 and 6 below derive the parameters of the Richardson reaction functions, a'_1, a'_3, b'_1, b'_3 of (3) and (4), as explicit functions of technical and timing factors in a simulated missile war. They also present equilibrium stocks (6) and (7) as explicit functions of these factors.

---

[3] It has been assumed that the fatigue coefficients, a_2 and b_2, are nonzero.

[4] For a discussion of general reaction functions see Intriligator (1964).

[5] Richardson equated the upward spiral with war outbreak, a rather contrived way to bring war outbreak into the model. In later work, e.g., Boulding (1962) and Intriligator (1964), it was noted that the upward spiral of missiles is halted at constraint levels, at which each side holds the maximum number of missiles consistent with its resources and institutions.

3. A Dynamic Model of a Missile War[6]

A basic premise of this paper is that decisions to acquire missiles must be based on how these missiles would be utilized in a simulated war. This simulated war can be summarized by the following differential equations and boundary conditions, which constitute a dynamic model of a missile war:

$$\dot{M}_A = -\alpha M_A - \beta'\beta M_B f_B, \qquad M_A(0) = M_A^0, \tag{8}$$

$$\dot{M}_B = -\beta M_B - \alpha'\alpha M_A f_A, \qquad M_B(0) = M_B^0, \tag{9}$$

$$\dot{C}_A = (1 - \beta')\beta M_B v_B, \qquad C_A(0) = 0, \tag{10}$$

$$\dot{C}_B = (1 - \alpha')\alpha M_A v_A, \qquad C_B(0) = 0. \tag{11}$$

Here, as before, $M_A(t)$ and $M_B(t)$ are missile stocks in countries A and B, respectively, at time t, while $C_A(t)$ and $C_B(t)$ are casualties in A and B, respectively, at time t. The war starts at time $t = 0$, and each country enters the war with certain initial stocks of missiles, M_A^0 and M_B^0, and zero casualties. Country A fires its missiles at the rate $\alpha(t)$, so $-\alpha M_A$ in (8) represents the reduction in A missiles due to the firing decisions of A. Missiles can be targeted at either enemy cities (counterforce) or enemy cities (countervalue), and, assuming A uses the counterforce proportion $\alpha'(t)$, of the αM_A missiles launched by A, $\alpha'\alpha M_A$ are targeted at B missiles while $(1 - \alpha')\alpha M_A$ are targeted at B cities. The $\alpha'\alpha M_A$ counterforce missiles launched by A destroy $\alpha'\alpha M_A f_A$ of the B missiles, where $f_A(t)$ is the number of B missiles destroyed by one A counterforce missile.[7] Thus, in (9), $\alpha'\alpha M_A f_A$ are the B missiles destroyed by A missiles. The $(1 - \alpha')\alpha M_A$ countervalue missiles launched by A inflict $(1 - \alpha')\alpha M_A v_A$ casualties in B, as indicated in (11). A similar analysis holds for country B decisions to launch missiles, $\beta(t)$, and to target enemy missiles, $\beta'(t)$.

The evolution of the war over time, summarized by (8)–(11), depends on initial missiles, M_A^0, M_B^0; strategic decisions regarding rates of fire, $\alpha(t)$, $\beta(t)$; strategic decisions regarding targeting, $\alpha'(t)$, $\beta'(t)$; the effectiveness of missiles against enemy missiles, $f_A(t)$, $f_B(t)$; and the effectiveness of missiles against enemy cities, $v_A(t)$, $v_B(t)$. For convenience, the time dependence of variables has been omitted in (8)–(11).

---

[6] See Intriligator (1967, 1968).

[7] The effectiveness of missiles against missiles depends on the yield and accuracy of the attacking missiles and the vulnerability of the attacked missile. In general,

$$f_A = \gamma_A \frac{W_A^{2/3} K_B^2}{I_A^2},$$

where W_A is the yield of an A missile, K_B is the vulnerability (lethal radius) of a B missile, I_A is the inaccuracy (circular error probable) of an A missile, and γ_A is a constant dependent on the reliability of A missiles and the degree of confidence sought in destroying B missiles. As any of these factors change, f_A changes, e.g., improved yield or accuracy of A missiles or hardening of B missiles.

4. Simulated War Outbreak

Of perhaps greatest relevance to defense planners concerned with the acquisition of missiles is the outbreak of war. The outbreak of a simulated war can, however, be analyzed using the dynamic model (8)–(11).

Assume A initiates the simulated war by attacking B. In order to reduce the retaliatory capability of B, country A initiates the war by choosing its maximum rate of fire, $\alpha = \bar{\alpha}$, selecting enemy missiles as targets, $\alpha' = 1$.[8] This first strike lasts a certain length of time, θ_A, the length of time before the attacked country B responds. Over this "window" time interval, $0 \leq t \leq \theta_A$, A attacks without a response from B, so $\beta = 0$. The window θ_A is assumed to be so short (e.g., of the order of minutes) that the effectiveness of A missiles against B missiles, v_A, is constant. Inserting $\alpha = \bar{\alpha}$, $\alpha' = 1$, and $\beta = 0$ into (8)–(11) and integrating these equations, at the end of the window interval

$$M_A(\theta_A) = M_A^0 \exp(-\bar{\alpha}\theta_A), \tag{12}$$

$$M_B(\theta_A) = M_B^0 - f_A[1 - \exp(-\bar{\alpha}\theta_A)]M_A^0. \tag{13}$$

Thus, during the first strike A expends $[1 - \exp(-\bar{\alpha}\theta_A)]M_A^0$ of its missiles to destroy $f_A[1 - \exp(-\bar{\alpha}\theta_A]M_A^0$ of the B missiles. The remaining B missiles can be used to retaliate against A.

In its retaliatory strike, it is assumed that B chooses its maximum rate, $\beta = \bar{\beta}$, and countervalue targets, $\beta' = 0$. Defense planners in A studying this simulated war, using "worst-case" analysis and believing the retaliatory strike might be countervalue, would make plans on this basis.[9] The retaliation period, from time $t = \theta_A$ to time $t = \theta_A + \psi_B$, is thus characterized by $\beta = \bar{\beta}$, $\beta' = 0$, and $\alpha = 0$.[10] The retaliation window ψ_B is assumed so short (e.g., of the order of minutes) that the effectiveness of B missiles against A cities, v_A, is constant. Integrating (8)–(11) with these assumptions as to strategy, the number of casualties in A at the end of the retaliatory strike is[11]

$$C_A(\theta_A + \psi_B) = v_B\{M_B^0 - f_A[1 - \exp(-\bar{\alpha}\theta_A)]M_A^0\}[1 - \exp(-\bar{\beta}\psi_B)]. \tag{14}$$

---

[8] The analysis can be easily modified to allow for any rate of fire α and any counterforce proportion α'. In Intriligator (1967), however, it was proved that in order to maximize a payoff function dependent on terminal missiles and casualties in both countries, the opening phase of the war would optimally involve the maximum rate of fire ($\alpha = \bar{\alpha}$) and counterforce targeting ($\alpha' = 1$). The closing phase would involve countervalue targeting and the minimum rate of fire. The middle phase could be either counterforce at the zero rate or countervalue at the maximum rate, depending on the effectiveness of missiles and the enemy strategy.

[9] As noted in the previous footnote, in Intriligator (1967) it was demonstrated that the middle phase of a missile war might be countervalue at the maximum rate, which is interpreted here as a potential counterforce retaliatory strike. The analysis can also be easily modified, however, to allow for any retaliatory rate of fire β and any retaliatory counterforce proportion β', rather than the assumed values of $\beta = \bar{\beta}$ and $\beta' = 0$.

[10] The analysis can be easily modified to allow for $\alpha = \bar{\alpha}$ during the retaliatory phase.

[11] Boundary conditions are matched at time θ_A.

This result of a simulated war and the corresponding symmetric result if B initiates, $C_B(\theta_B + \psi_A)$, will be utilized to derive Richardson-type reaction functions.

5. Strategy and Arms Races: Two Deterrers

Defense planners in the countries involved in an arms race are assumed to adopt one of two alternative objectives: the "deterring" objective and the "attacking" objective. Under the deterring objective, the country seeks to deter its opponent from striking first by threatening to inflict upon it an unacceptable level of damage on a potential second strike. Since the opponent is assumed to value population (and industrial capacity) more highly than military power, a deterrer would utilize a countervalue second strike directed against enemy cities. The alternative objective is the attacking objective under which the country seeks to disarm the opponent by destroying sufficient numbers of its missiles in a counterforce first strike such that the casualties inflicted on a potential retaliatory strike are reduced to acceptable levels.

Assuming A and B are both deterrers, each should have sufficient numbers of missiles to inflict unacceptable damage on the other. If country B believes the minimum unacceptable damage to A is $\bar{C}_A$ casualties, it should have enough missiles to inflict this number of casualties on a second strike. Equation (14), however, indicates the number of casualties B would inflict in exactly this situation for a simulated war. Setting this number equal to $\bar{C}_A$, solving for M_B^0, the required initial number of B missiles, and replacing M_B^0 by M_B and M_A^0 by M_A, since war could break out at any time and the M_A^0, M_B^0 levels are simply initial levels of missiles in the simulated war, yields

$$M_B = \{f_A[1 - \exp(-\bar{\alpha}\theta_A)]\}M_A + \left\{\frac{\bar{C}_A}{v_B[1 - \exp(-\bar{\beta}\psi_B)]}\right\}. \qquad (15)$$

This equation shows the number of B missiles needed to deter A as a function of the number of A missiles. It is, in fact, the Richardson reaction function for country B, equation (4), for which the (normalized) Richardson reaction coefficients are

$$b_1' = f_A[1 - \exp(-\bar{\alpha}\theta_A)], \qquad (16)$$

$$b_3' = \frac{\bar{C}_A}{v_B[1 - \exp(-\bar{\beta}\psi_B)]}. \qquad (17)$$

The normalized defense term b_1' is positive and given in (16) as an explicit function of the effectiveness of A missiles against B missiles, f_A; the maximum rate of fire of the A missiles, $\bar{\alpha}$; and the window time interval for A, θ_A. Since f_A is likely to be less than unity (i.e., it takes more than one missile to destroy a B missile) and $[1 - \exp(-\bar{\alpha}\theta_A)]$ is always less than unity, the product of these terms, which is the normalized defense coefficient of the Richardson model, is generally less than unity.

The normalized grievance term, b'_3, is positive and given in (17) as an explicit function of the number of A casualties required for B to deter A, $\bar{C}_A$; the effectiveness of B missiles against A cities, v_B; the maximum rate of fire of the B missiles, $\bar{\beta}$; and the retaliation window time interval, ψ_B.

A similar analysis and corresponding notation is used if B attacks, so that the number of A missiles required to deter B is

$$M_A = \{f_B[1 - \exp(-\bar{\beta}\theta_B)]\}M_B + \left\{\frac{\bar{C}_B}{v_A[1 - \exp(1 - \bar{\alpha}\psi_A)]}\right\}. \tag{18}$$

This is again the reaction curve for a normalized Richardson model, where, in (3), the coefficients are

$$a'_1 = f_B[1 - \exp(-\bar{\beta}\theta_B)], \tag{19}$$

$$a'_3 = \frac{\bar{C}_B}{v_A[1 - \exp(-\bar{\alpha}\psi_A)]}. \tag{20}$$

Since the grievance terms are positive, a stable equilibrium exists if, as in (5), the product of these slope coefficients is less than unity,

$$a'_1 b'_1 = f_A f_B[1 - \exp(-\bar{\alpha}\theta_A)][1 - \exp(-\bar{\beta}\theta_B)] < 1. \tag{21}$$

This condition is likely to be met since f_A and f_B are probably less than unity and the other two terms are always less than unity. It should be noted that the "hardness" condition that more than one missile is required to destroy an enemy missile ($f_A, f_B < 1$) is sufficient but not necessary for stability. It should also be noted that the stability condition is independent of $\bar{C}_A$, $\bar{C}_B$, v_A, and v_B. In particular, the levels of minimum unacceptable casualties do not influence the stability of equilibrium. In the case of two deterrers, *the equilibrium is always stable provided condition (21) is met.*

Assuming (21) holds, a stable equilibrium exists, and solving as in (6) and (7) yields

$$M_A^E = \frac{\{f_B[1 - \exp(-\bar{\beta}\theta_B)]\bar{C}_A\}/\{v_B[1 - \exp(-\bar{\beta}\psi_B)]\} + \bar{C}_B/\{v_A[1 - \exp(-\bar{\alpha}\psi_A)]\}}{1 - f_A f_B[1 - \exp(-\bar{\alpha}\theta_A)][1 - \exp(-\bar{\beta}\theta_B)]}, \tag{22}$$

$$M_B^E = \frac{\{f_A[1 - \exp(-\bar{\alpha}\theta_A)]\bar{C}_B\}/\{v_A[1 - \exp(-\bar{\alpha}\psi_A)]\} + \bar{C}_A/\{v_B[1 - \exp(-\bar{\beta}\psi_B)]\}}{1 - f_A f_B[1 - \exp(-\bar{\alpha}\theta_A)][1 - \exp(-\bar{\beta}\theta_B)]}. \tag{23}$$

These are the levels of missiles to which the system will tend, assuming the stability condition is met. Over time the technical and timing parameters in these equations may change, but at any time the system will tend to the levels determined by the parameters at that time.

Equations (22) and (23) give the explicit dependence of equilibrium missile stocks on technological factors and decisions, in particular, decisions involving missile hardening, accuracy, and yield (affecting f_A and f_B); civil defense (affecting v_A and v_B); warning time (affecting θ_A and θ_B); and command and control (affecting $\bar{\alpha}$ and $\bar{\beta}$). To give a numerical example, suppose a symmetric situation where 40 million casualties will deter the opponent, one missile can inflict 250,000 casualties, it takes two missiles to destroy one enemy missile, the rate of fire is 10 percent per minute, the first strike window is 15 minutes, and the second strike retaliatory window is 10 minutes, that is,

$$\begin{aligned} \bar{C}_A &= \bar{C}_B = 40 \text{ million (casualties)}, \\ v_A &= v_B = 250{,}000 \text{ (casualties per missile)}, \\ f_A &= f_B = 0.5 \text{ (missiles per missile)}, \\ \bar{\alpha} &= \bar{\beta} = 0.10 \text{ (percent per minute)}, \\ \theta_A &= \theta_B = 15 \text{ (minutes)}, \\ \psi_A &= \psi_B = 10 \text{ (minutes)}. \end{aligned} \tag{24}$$

The equilibrium stock of missiles held by either side is then approximately 414 missiles, that is,

$$M_A^E = M_B^E = 414. \tag{25}$$

In a simulated war, if A attacked B, during the first 15 minutes of the war A could launch 322 missiles against B missiles, destroying 161 B missiles. The remaining 253 missiles would be launched at the maximum rate over the next 10 minutes, during which 160 missiles would be launched at A cities. These missiles could inflict 40 million casualties in A.

Table 1 shows the sensitivity of the equilibrium missile stocks to changes in the A parameters, where, in each case, all other parameters are given in (24). Thus, for example, an increase in the minimum unacceptable level of casualties in A, $\bar{C}_A$, to 60 million, all other parameters held constant, would increase both equilibrium levels of missiles but would reasonably increase the equilibrium level for B (to 563) by more than that for A (to 472). Decreasing the countervalue effectiveness of A missiles to $v_A = 100{,}000$, for example, by installation of civil defense measures in B, would also increase the equilibrium levels of missiles for both countries but for A (to 861) by more than for B (to 588).

Table 1 and the explicit equations (22) and (23) could be used to identify tradeoffs among the technical and timing parameters. For example, given the other parameters, a decrease of $\bar{C}_A$ to 20 million has approx-

TABLE 1

SENSITIVITY OF EQUILIBRIUM MISSILE STOCKS TO CHANGES IN PARAMETERS*

Parameter for A	M_A^E	M_B^E
$\bar{C}_A$ = 60 million	472	563
$\bar{C}_A$ = 20 million	356	269
v_A = 500,000................	265	356
v_A = 100,000................	861	588
f_A = 0.2	374	311
f_A = 0.8	463	541
$\bar{\alpha}$ = 0.05	563	402
$\bar{\alpha}$ = 0.2	347	418
θ_A = 10	401	380
θ_A = 20	422	436
ψ_A = 5......................	595	484
ψ_A = 15....................	358	392

*In each case the other parameters are given in (24).

imately the same effect on M_A^E as an increase in $\bar{\alpha}$ to 0.2 or an increase in ψ_A to 15. Thus, an increase in the A retaliatory period from 10 to 15 minutes has the same effect on the equilibrium number of A missiles (but not on the equilibrium number of B missiles) as a decrease in the minimum unacceptable A casualties from 40 million to 20 million casualties.

Table 1 also illustrates the opposite effects of ABM (antiballistic missile) systems depending on whether they protect cities or missiles. ABM defense of cities in B reduces v_A and hence, as shown in the table, increases the equilibrium levels of missiles in both countries. ABM defense of missiles in B, however, reduces f_A and therefore *decreases* the equilibrium levels of missiles in both countries.

6. Strategy and Arms Races: Attackers and Asymmetric Cases

The stable equilibrium obtained in the last section assumed two deterrers. Consider now the case of two attackers. If A is the attacker it should have sufficient numbers of missiles to so reduce the number of B missiles on its first strike in a simulated war that B could not inflict unacceptable damage on its retaliatory strike. If the maximum acceptable damage to A is considered *by* A to be $\hat{C}_A$ casualties, then country A should have eliminated, on its first strike, enough B missiles so that B could at most inflict $\hat{C}_A$ casualties on its retaliatory strike. Again using equation (14) but now solving for M_A^0, the required number of A missiles, and replacing M_A^0 by M_A and by M_B^0 by M_B, as before, yields

$$M_A = \left\{\frac{1}{f_A[1 - \exp(-\bar{\alpha}\theta_A)]}\right\} M_B - \left\{\frac{\hat{C}_A}{f_A[1 - \exp(-\bar{\alpha}\theta_A)]v_B[1 - \exp(-\bar{\beta}\psi_B)]}\right\}. \tag{26}$$

This equation indicates how many missiles A needs, as a function of the number of B missiles, in order to be able to attack B. Again, the equation is simply a Richardson reaction function for A, as in equation (3). The normalized defense term depends on f_A, $\bar{\alpha}$, and θ_A, as shown in (26), and is generally *greater* than unity since it is simply the reciprocal of the normalized defense term in (16). The normalized grievance term depends on $\hat{C}_A$, f_A, $\bar{\alpha}$, θ_A, v_B, $\bar{\beta}$, and ψ_B, as shown in (26), and is negative. A similar analysis holds for the Richardson reaction function for B if B is an attacker. In this case of two attackers, the stability condition is not met but the grievance terms are negative. There therefore exists an equilibrium level of missiles for each side, but the equilibrium is unstable. The equilibrium levels of missiles are the same as those given as in (22) and (23), where $\bar{C}_A$ is replaced by $\hat{C}_A$ and $\bar{C}_B$ is replaced by $\hat{C}_B$. Since the equilibrium is unstable, slight movements above the equilibrium configuration will lead to further upward movements as each country seeks to obtain enough missiles to attack the other, while movements below the equilibrium will lead to further downward movements as each country finds it can reduce its stock of missiles and still have enough to attack the other.

So far, only symmetric cases have been treated where both A and B are deterrers, leading to a stable equilibrium, and where both are attackers, leading to an unstable equilibrium. The asymmetric case is where, by suitable relabeling, A seeks to attack B while B seeks to deter A. The reaction functions are (26) and (15), which are *parallel* lines in the M_A, M_B plane, implying that no equilibrium exists. The outcome in this case depends on the relation between $\bar{C}_A$ and $\hat{C}_A$, where $\bar{C}_A$ is the estimate by B of the minimum number of casualties which A would find unacceptable, and $\hat{C}_A$ is the estimate by A of the maximum number of casualties it would find acceptable. It is reasonable to assume that $\bar{C} \geq \hat{C}_A$, since both sides are likely to use a safety margin, boosting $\bar{C}_A$ and reducing $\hat{C}_A$. In this case there exists no equilibrium, and the outcome would be an upward spiral of missiles as A seeks to acquire missiles to attack B and, simultaneously, B seeks to acquire missiles to deter A. The spiral would end at certain constraint levels.[12] Alternatively, A might switch to a deterring strategy.

7. Arms Control and War Initiation

The preceding analysis of strategy and arms races establishes a basis for analyzing arms control and war initiation. While in previous sections the deterring and attacking objectives are treated as distinct alternatives, in the context of war initiation both objectives must be treated. Each country must consider both deterring and attacking the opponent, since the outbreak of an actual war depends both on the incentive to attack and on the failure of deterrence.

[12] See n. 5 above, Boulding (1962), and Intriligator (1964).

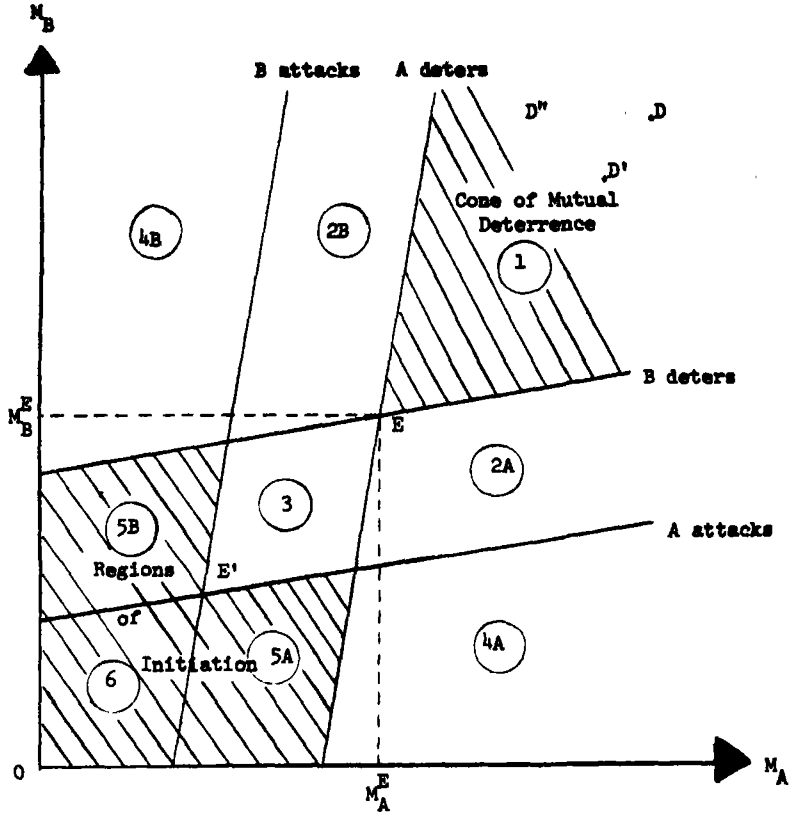

FIG. 1.—Deterring and attacking reaction curves for both countries

Reaction curves for each country acting as deterrer or attacker are shown geometrically in the missile plane M_A, M_B in figure 1.[13] The line marked "*A* deters" shows the minimum number of *A* missiles needed to deter *B* as given in equation (18), its slope being $\{f_B[1 - \exp(-\bar{\beta}\theta_B)]\}^{-1}$, which is assumed larger than unity. The line marked "*B* deters" similarly shows the minimum number of *B* missiles needed for *B* to deter *A*, as given in equation (15), its slope being $f_A[1 - \exp(-\bar{\alpha}\theta_A)]$, which is assumed to be smaller than unity. The line marked "*A* attacks" shows the minimum number of *A* missiles needed for *A* to attack *B*, as given in equation (26), its slope being $f_A[1 - \exp(-\bar{\alpha}\theta_A)]$, so it is parallel to the "*B* deters" line. Finally, the line marked "*B* attacks" shows the minimum number of *B* missiles needed for *B* to attack *A*, its slope being $\{f_B[1 - \exp(-\bar{\beta}\theta_B)]\}^{-1}$, so it is parallel to the "*A* deters" line.

The equilibrium for two deterrers is shown in figure 1 as the point *E*, where the equilibrium stocks of missiles are defined by equations (22) and

---

[13] For related diagrams, see Kybal (1960) and Beaufre (1965).

(23). This is a stable equilibrium. By contrast, the equilibrium for two attackers at E' is unstable. The asymmetric case where one side is a deterrer and the other an attacker is represented by the parallel reaction curves, such as "A attacks" and "B deters." This case, involving no equilibrium, leads to an upward spiral of missiles.

The four reaction curves divide the missile plane into nine regions, as shown in figure 1, and each of these regions can be interpreted in the analysis of war initiation.

Region 1 is the cone of *mutual deterrence*, where each side holds sufficient numbers of missiles to deter the other and neither side can attack the other. All points in the shaded cone area are stable against war outbreak in that they represent situations of mutual assured destruction; neither country has enough missiles to be able to destroy sufficient enemy missiles on a first strike so that the damage the other side could inflict on a retaliatory strike would fall within an "acceptable" level. The minimal deterrent missile stocks at E are given in equations (22) and (23), and their sensitivities to changes in parameters are given in table 1.[14]

Regions 2A and 2B are those of *advantage to one side*. For example, in 2A country B does not have enough missiles to deter A and, while A has enough missiles to deter B, it does not have enough to attack B.

Region 3 is one of *jittery deterrence* in which neither side can attack the other but, in addition, neither side can deter the other.

Regions 4A and 4B are those of *dangerous power on one side*. For example, in 4A country A deters and can attack B while country B can neither deter nor attack A.

Regions 5A and 5B are those of *forced initiation by one side*. For example, in 5A country A has enough missiles to attack B but not enough to deter B while B has insufficient numbers of missiles to either deter or attack A. The result is that A is forced to initiate, since it has enough missiles for a first strike but not enough for a second strike. Only by preempting can A eliminate the fear of an attack by B which A is unable to deter.

Region 6 is one of *forced initiation by both sides* and is clearly the most explosive of the nine regions. In this region, either country can attack the other but neither can deter the other. The "reciprocal fear of surprise attack" based on the tremendous advantage in striking first forces both sides to initiate, each trying to preempt the attack of the other.[15]

Regions of asymmetry, where one side has a large stockpile and the other has relatively few missiles, can be stable against war outbreak if the more powerful country has peaceful intentions. Even with the most peaceful of intentions, however, if one or both countries can attack the other but cannot deter it, the premium for a first strike will make the situation highly explosive. Thus, regions 5*A*, 5*B*, and 6 are shown as regions of war initiation.

---

[14] See also Szilard (1964).

[15] See Wohlstetter (1959).

This analysis has several implications for arms control and disarmament. Complete disarmament is interpreted geometrically as a movement to the origin in figure 1, where $M_A = M_B = 0$. The process of disarmament can lead to highly explosive combinations of missile stocks, which could lead to war outbreak. Even if the process of disarming does not lead to war, the disarmed state itself is explosive, since, from it, either country could acquire a minimal stock of missiles which would enable it to attack the other with impunity. The danger of war outbreak may in fact be greatest if one side detected that the other was starting to rearm, since in these cases the missile configurations move through explosive regions in a crisis atmosphere.

Unilateral disarmament, interpreted geometrically in figure 1 as a move to zero missiles by one country, may be safe in not leading to a region of initiation provided the other country has enough missiles. Thus, for example, if *B* holds M_B^E missiles it may be safe for *A* to unilaterally disarm, trusting in *B*'s intentions. If *B* had significantly fewer missiles, however, a region of initiation would be entered, forcing *B* to initiate.

Arms control is interpreted geometrically as movements to more desirable points in the missile plane. One possible interpretation of strategic arms limitation talks is that, due to historical factors (e.g., past values of effectiveness ratios f_A, f_B, v_A, v_B; of rates of fire $\bar{\alpha}, \bar{\beta}$; and of window time intervals $\theta_A, \theta_B, \psi_A, \psi_B$), the two countries are quite far up in the cone of mutual deterrence and could therefore reasonably agree jointly to disarm toward equilibrium levels, at all times staying within the cone of mutual deterrence. Of course, each side may reasonably insist, because of uncertainty, on more missiles than at the equilibrium, but there still may be scope for arms limitation agreements. Bilateral arms limitation agreements, moving the configuration of missiles from a point high up in the cone, such as *D*, to a point closer to the origin for both but also in the cone, such as *D*′, might be desirable. Such a move would reduce missile stocks and hence the cost of maintaining missiles, the probability of accidental war, and the severity of war should it break out, all without jeopardizing mutual deterrence. Even unilateral arms reductions, such as from *D* to *D*″, are possible without upsetting the stability of mutual deterrence, provided the reduction is not large enough to move the configuration of missile stocks out of the cone. Arms control may also, however, require *increases* in missile stocks. For example, a movement from a region of initiation to the cone of mutual deterrence by both countries acquiring adequate numbers of missiles could stabilize a potentially explosive situation. Arms control measures must, however, according to the analysis presented here, take into account present and future technical and timing factors which influence the intercepts and slopes of the reaction curves of figure 1. Technological developments could very well change the equilibrium levels of missiles and thereby calm an explosive situation or upset a peaceful situation.

8. Conclusions

This paper has shown how the coefficients of Richardson-type reaction functions for levels of missile stock can be derived on the basis of strategic objectives (deterring or attacking objectives), technical considerations (the effectiveness of missiles against both missiles and cities and the maximum rates of fire available), and timing factors (the window time intervals for a first strike and for a retaliatory second strike).[16] If both countries are deterrers there generally exists a stable equilibrium of mutual assured destruction for which each side holds sufficient missiles to deter the other by threatening to inflict unacceptable damage on it in a retaliatory second strike. If both countries are attackers, seeking to destroy sufficient numbers of enemy missiles on a first strike to ensure that the damage in a retaliatory strike is held to acceptable levels, then there generally also exists an equilibrium level of missile stocks, but the equilibrium is unstable. If one side is a deterrer and the other an attacker, then no equilibrium exists and the outcome is probably an upward spiral in levels of missile stocks as the attacker seeks to obtain enough missiles to attack, while the deterrer simultaneously seeks to obtain enough missiles to deter. The Richardson-type reaction functions imply the existence of a cone of mutual deterrence and also regions of war initiation in the missile plane, defined by the numbers of missiles in both countries. Unilateral and bilateral arms control and disarmament measures are analyzed in terms of movements in this plane.

References

Abelson, R. P. "A 'Derivation' of Richardson's Equations." *J. Conflict Resolution* 7 (1963): 13–15.

Beaufre, A. *Deterrence and Strategy*. Translated from the French by R. H. Barry. New York: Praeger, 1965.

Boulding, K. *Conflict and Defense: A General Theory*. New York: Harper, 1962.

Burns, A. L. "A Graphical Approach to Some Problems of the Arms Race." *J. Conflict Resolution* 3 (1959): 326–42.

Intriligator, M. D. "Some Simple Models of Arms Races." *General Systems* 9 (1964): 143–47.

———. *Strategy in a Missile War: Targets and Rates of Fire*. Security Studies Project, University of California. Los Angeles: Univ. California, 1967.

---

[16] For another approach to the derivation of the coefficients of the Richardson model, see Abelson (1963), where the coefficients are based on "sensitivity to" and "forgiveness of" the actions (weapon accumulations) of the opponent.

———. "The Debate over Missile Strategy: Targets and Rates of Fire." *Orbis* 11 (1968): 1138–59.

Kent, G. A. *On the Interaction of Opposing Forces under Possible Arms Agreements.* Cambridge, Mass.: Center Internat. Affairs, Harvard Univ., 1963.

Kybal, D. "Remarks." In *Proceedings* of the National Strategy Seminar, Asilomar, Calif., April 24–30, 1960.

McGuire, M. C. *Secrecy and the Arms Race.* Cambridge, Mass.: Harvard Univ. Press, 1965.

Pitman, G. R., Jr. *Arms Races and Stable Deterrence.* Security Studies Project, University of California. Los Angeles: Univ. California, 1969.

Rapoport, A. "Lewis F. Richardson's Mathematical Theory of War." *J. Conflict Resolution* 1 (1957): 249–304.

———. *Fights, Games, and Debates.* Ann Arbor: Univ. Michigan Press, 1960.

Richardson, L. F. "Generalized Foreign Politics." *British J. Psychology Monographs Suppl.*, vol. 23 (1939).

———. *Arms and Insecurity.* Pittsburgh: Boxwood, 1960.

Saaty, T. L. *Mathematical Models of Arms Control and Disarmament.* New York: Wiley, 1968.

Schelling, T. C. *Arms and Influence.* New Haven, Conn.: Yale Univ. Press, 1966.

Szilard, L. "'Minimal Deterrent' vs. Saturation Parity." *Bull. Atomic Sci.* 20 (1964): 6–12.

Wohlstetter, A. "The Delicate Balance of Terror." *Foreign Affairs* 37 (1959): 211–34.

Unemployment, Underemployment, and Optimal Job Search

B. Curtis Eaton and Philip A. Neher

University of British Columbia

A three-sector general equilibrium model is employed to analyze the consequences of changes in aggregate demand, union power, workers' time horizons, and labor force size on "unionized" industries, industries with competitive labor markets, and unemployment. Workers are "voluntarily" unemployed while searching for jobs, attempting to arbitrage a wage gap created by partial "unionization." Many reduced-form elasticities are ambiguous in algebraic sign. For example, decreased aggregate demand may increase employment in industries with competitive labor markets. However, a larger and younger work force with enhanced aspirations for higher-paying jobs will have unambiguously higher unemployment.

> In all those occupations which the dismissed workers take up, their productivity is less than in the occupations they have left. For if it were not so, they would have engaged in them already. [JOAN ROBINSON, 1936]
>
> I'm really a university professor. Recently, I've been stringing beads for a living. [ANONYMOUS, 1974]

I. Introduction

Neoclassical economics teaches that factors of production will search out and find occupations of highest reward. Yet we observe people in occupations of last resort "selling match boxes in the Strand, cutting brushwood in the jungles, digging potatoes on allotments" (Robinson 1936, p. 226). Failing to find jobs for which they are suited by training

---

We are indebted to members of the University of British Columbia Labour Workshop for valuable comments. In addition, we owe a great debt to an anonymous referee who has twice offered invaluable comments which have led us to a final product that is very much better than, and substantially different from, our original effort.

[*Journal of Political Economy*, 1975, vol. 83, no. 2]

© 1975 by The University of Chicago. All rights reserved.

and temperament, they disappear into the statistical never-never land of underemployment where rewards are meager but available. University graduates selling underground newspapers on campuses, flowers on street corners, narcotics on city streets, or encyclopedias from door to door illustrate underemployment of the young and educated. Even the "hard-core" employed are affected. The laid-off tool and die maker reappears pumping gasoline or sharpening knives from door to door. The marine architect returns to the family farm and feeds the chickens.

Underemployment has hardly been ignored in the literature. But since the classic articles of the 1930s, it has been chiefly studied in the context of underdeveloped or semideveloped economies (Rosenstein-Rodan 1943; Lewis 1955; Fei and Ranis 1964; Todaro 1969; Harris and Todaro 1970). Yet, if the Great Depression gave birth to the concept in the framework of an industrialized society, our current circumstances require that we return for another look.

Recently, H. G. Johnson and P. Mieszkowski and E. Diewert showed that a properly constrained general equilibrium model can be used to investigate the consequences of successful unionization of part of the labor force (Johnson and Mieszkowski 1970; Diewert 1972). Union labor works in the relatively high-paid "general run of industries," while, in our interpretation, the rest of the labor force is "underemployed." But a wage gap is an invitation for the underpaid to seek out high-paid job opportunities as they become available due to natural turnover.

In this view, unemployment and underemployment are *concomitant* consequences of "union power." Union power gives rise to underemployment and wage dispersion. Wage dispersion gives rise to unemployment as workers try to arbitrage the wage gap. If people must quit low-paid work to search for high-paid work, it is natural to turn to a job search model in the tradition of M. P. Todaro (1969), C. C. Holt (1970), D. T. Mortensen (1970), and A. C. Harberger (1971) to explain unemployment.

Our contribution is twofold. First, we refine the usual job search model by articulating an explicit aggregation procedure to derive an unemployment function. Second, the unemployment function is embedded in a general equilibrium framework.

The concept of equilibrium in an economy with institutionally generated wage differentials requires careful interpretation. Differentials may reflect real differences in skill, training, cost of living, and so on. In this case, the institution (union) simply validates a free-market solution. We have in mind a true "contrived scarcity" created by people banding together to create a monopoly in the sale of their labor. Our "equilibrium" recognizes that such monopolies can be successful over the period of time relevant to the analysis. If pressed to justify our assumption of effective institutional power, we should probably proceed along lines suggested by C. C. von Weizsäcker in his Wicksell Lectures (Weizsäcker 1972). We do not pursue the point here. We simply assume that institutionally generated wage differentials can persist.

But institutions are sluggish in response to changing circumstances. Three-year union-management contracts, for example, are not uncommon in Canada. Hence, our idea of equilibrium includes the concept of an institutional *rigidity* which is effective over the relevant period of analysis. In short, we assume that institutional (union) power to create wage differentials exists but that it can be temporarily diluted from time to time because of a sluggish response to unexpected changes in the economy at large.

We shall first set out a short-run, general equilibrium, comparative statics model which allows for unemployed labor searching for high-wage jobs. The model is then solved for the reduced-form elasticities of chief concern. We use these results to assess demand management policies and to trace out the consequences of changes in demographic variables, social attitudes, and the degree of union power. We conclude with conclusions.

II. The Model

There are three sectors: the X industry (characterized by labor market rigidities), the Y sector (marked by flexible wages and ease of entry), and an unemployment sector.

A. The X Industry

We include here all activities where access to jobs is limited (in the sense that an applicant cannot enhance his employment prospects by offering his services at a wage lower than the prevailing one) due to institutional constraints. We have in mind Harberger's "protected industries" or Robinson's "general run of industries" (Robinson 1936; Harberger 1971).

The short-run production function is conditioned by the quantity of nonhuman resources already committed to that sector, so that output (X) depends on the labor input (L_X):

$$X = F(L_X), \qquad F' > 0, F'' < 0. \tag{1}$$

Labor is paid the value of its marginal product:

$$w_X = F'p, \tag{2}$$

where p is the price of X.

Institutional rigidities are embodied in the supply of labor function. We characterize these rigidities as two polar cases. First, a rigid supply of labor is maintained by restricting entry through unions, professional associations, and the like:

$$L_X = \bar{L}_X. \tag{3a}$$

We take it that $\bar{L}_X$ is set to obtain a wage in excess of the free-market wage. Alternatively, a rigid wage is set above the free-market wage so that the supply of labor is perfectly elastic over the relevant range of employment:

$$w_X = \bar{w}_X. \tag{3b}$$

The quantity of X demanded is a function of relative prices and real income. Money income (equals aggregate demand, D) is, for simplicity, assumed proportional (by V) to the exogenous nominal money supply $(\bar{M})$, $D = V\bar{M}$. For notational convenience, we shall simply let aggregate demand be exogenous $(D = \bar{D})$. Then, normalizing on the price of $Y(q)$, X demanded depends on its relative price (p/q) and on real income, or real aggregate demand, $(\bar{D}/q)$:

$$X = X(p/q, \bar{D}/q), \qquad X_1 < 0, X_2 > 0. \tag{4}$$

B. The Y Industry

The short-run production function for Y is

$$Y = G(L_Y), \qquad G' > 0, G'' < 0. \tag{5}$$

The wage in Y is the value of L_Y's marginal product:

$$w_Y = G'q, \tag{6}$$

where q is the price of Y. Labor supplied to Y is the residual of all labor $(\bar{L})$ which is not employed in X (L_X) or unemployed (L_U):

$$L_Y = \bar{L} - L_X - L_U. \tag{7}$$

The quantity of Y demanded is given by the budget constraint $\bar{D} = Xp + Yq$, or

$$Y = \frac{\bar{D}}{q} - \frac{Xp}{q}. \tag{8}$$

Note that (8) can be written $Y = Y(p/q, \bar{D}/q)$. We assume that $Y_1 > 0$.

C. The Unemployment Sector

Employment in the high-wage industry is preferred to unemployment and to employment in the low-wage industry (underemployment). We assume that in order to obtain a job in the X industry, an individual must specialize in information generation; he must be unemployed in search of a job in X. (The weaker assumption that the probability of obtaining a job in the X industry is greater for the unemployed than for the underemployed would not change the qualitative nature of our results.)

Given a turnover rate, t, in industry X, which is demographically determined by retirements and deaths, the number of job vacancies in industry X is tL_X per period. Assuming a random process of matching between unemployed persons and job vacancies, the probability that an unemployed person will find a job in X in one period is tL_X/L_U. This implies that an individual must expect to experience approximately $L_U/(tL_X)$ periods of unemployment before obtaining a job in industry X. Since entry into the underemployed (Y) sector is assumed to be free, the expected value of earnings foregone in search of a job in X is

$$w_Y \int_0^u e^{-rt}\, dt,$$

where u, the expected duration of unemployment, is $L_U/(tL_X)$ and r is the individual's rate of discount.

On the other hand, if the individual expects the wage differential to persist indefinitely, he will expect to gain

$$(w_X - w_Y) \int_u^K e^{-rt}\, dt$$

from a decision to be unemployed in search of a job in X, where K is the number of periods left in the individual's working life. Thus, the individual will choose to be unemployed (underemployed) if

$$(w_X - w_Y) \int_u^K e^{-rt}\, dt \ (\gtrless)\ w_Y \int_0^u e^{-rt}\, dt.$$

Evaluating the integrals and rearranging this decision rule can be written as

$$e^{-ru}\ (\gtrless)\ \frac{[w_X e^{-rK} - w_Y(1 - e^{-rK})]}{w_X}.$$

Taking logs and rearranging, we have

$$L_U\ (\gtrless)\ \frac{tL_X}{r}\, \{\log w_X - \log [w_X e^{-rK} + w_Y(1 - e^{-rK})]\}.$$

Up to this point, our model of the individual's unemployment-underemployment decision is substantially the same as that developed by Todaro for analysis of the rural-urban migration decision (Todaro 1969). His rural sector is our Y industry, his urban sector our X industry. Todaro postulates dependence of the rate of growth of the urban labor force on the difference in discounted-income prospects between the two regions.[1] Todaro's formulation, which assumes that the planning horizon, the cost of migration, and the rate of discount are the same for all rural workers, shares a problem common to most migration and job search models: it implies that all (or none) of the rural labor force ought to migrate. The introduction of individual differences with respect to discount rates and age, as is done in this paper, or the introduction of some random selection procedure enables one to skirt this problem.[2] There remain the problems (which are also common to most migration and job search models) of stability of the job search function over time and of aggregation.

In any disequilibrium situation which is working itself out, one is aggregating some decision rule over a changing population. This immediately implies that *one ought not to expect such relationships to be stable over time.* We turn first to the problem of aggregating the decision rule at a point in time. For an infinite horizon (letting $K \rightarrow \infty$) the individual decision rule reduces to

$$L_U \ (\gtrless) \ \frac{tL_X}{v} \log \left(\frac{w_X}{w_Y} \right),$$

where v is the discount rate appropriate to an infinite horizon. In the aggregation procedure and subsequent analysis, this inequality is easier to work with; however, it apparently eliminates individual differences with respect to the working horizon, and we are interested in analyzing the comparative statics with respect to changing time horizons (Sec. IV*C*). If we can somehow incorporate both the individual's discount rate and his time horizon in a pseudo discount rate over an infinite horizon, we can make use of the simplifications offered by the infinite horizon while retaining the ability to analyze the implications of changing time horizons. Letting v be the pseudo discount rate over an infinite horizon, we require that v be a function of the individual's discount rate, r, and his time horizon, K. This we obtain by equating the right-hand side of the two previous expressions and solving for v:

$$v = r \left\{ \frac{\log w_X - \log w_Y}{\log w_X - \log \left[w_Y + e^{-rK}(w_X - w_Y) \right]} \right\}.$$

Note that $\partial v / \partial r > 0$ and $\partial v / \partial K < 0$.

[1] Zarembka rightly notes that Todaro's model is more suited to the analysis of changes in the rural labor force (migration from rural to urban) than to the direct analysis of changes in the urban labor force (Zarembka 1970).

[2] Mortensen, for example, introduces randomness in the frequency with which employees consider the search decision (Mortensen 1970).

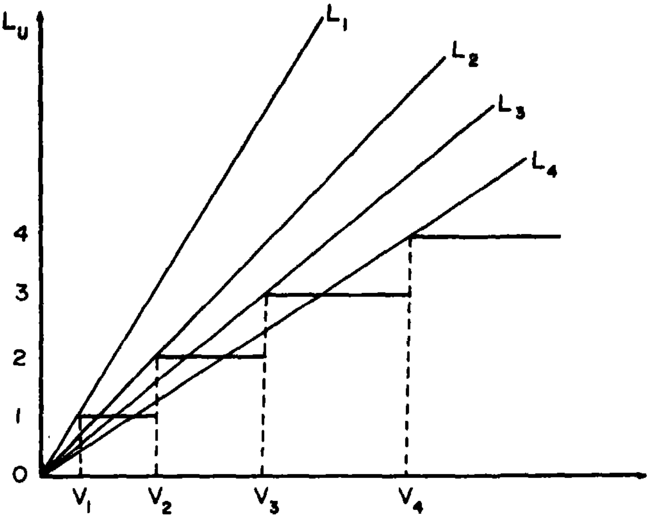

FIG. 1

For the jth individual, consider the locus of points

$$L_U = \frac{tL_X}{v_j} \log\left(\frac{w_X}{w_Y}\right) = JC_j,$$

where $C_j = 1/v_j$ and $J = tL_X \log(w_X/w_Y)$. Plot this locus in the (J, L_U) space. Our decision rule tells us that, for points below and to the right of the line, the jth individual will choose to be unemployed. For points above and to the left, he will choose to be underemployed. Several such loci are plotted in figure 1. Individual 1 in the figure is the least averse to unemployment. For J in the interval $[0, V_1]$, he will choose to be employed in industry Y. But in the interval $[V_1, V_2]$, he will choose to be unemployed while individuals 2, 3, and 4 will not. Accordingly, the aggregate unemployment function is 0 in the interval $[0, V_1]$, 1 in the interval $[V_1, V_2]$, 2 in the interval $[V_2, V_3]$, etc. Or, helping ourselves to a continuous approximation,

$$L_U = U(J), \qquad U' > 0, U'' < 0. \tag{9}$$

It is clear from the aggregation procedure that the second derivative of L_U with respect to J must be negative. This reflects the fact that the individuals who are least averse to unemployment are the first to become unemployed.

In each period, tL_X job vacancies occur in X, and, accordingly, tL_X of the unemployed obtain employment in X. And in each period all individuals in the Y and U sectors grow older, some die, and others retire. If the unemployment function is to be stable over time, in each period enough new individuals must enter the labor force to offset those who have taken employment in X plus those individuals in U and Y who have died or retired. Further, the entry of new individuals plus the aging of individuals in the U and Y sectors must be such as to maintain the distribution of the C_j's.

This completes the specification of the model. There are nine equations and nine unknowns (w_X, w_Y, p, q, L_X, L_Y, X, Y, L_U), with exogenous variables $\bar{w}_X$ or $\bar{L}_X$, $\bar{D}$, and L. The reduced-form (partial) elasticities are derived in the next section.

III. General Equilibrium

We wish to know how each of the endogenous variables, particularly unemployment, respond when an arbitrary change in one of the exogenous variables is imposed. Total (logarithmic) differentiation of equations (1)–(9) expresses the structural equations in terms of proportional changes. Cramér's rule could then be employed to render the system in the desired reduced form. Instead, it is more instructive to "solve" the model by combining equations market by market until only one endogenous variable appears in each equation. Our strategy is first to combine equations relating to the product markets (parameterizing momentarily on labor market variables), then to combine equations relating to the labor markets (parameterizing momentarily on product market variables), and finally to bring together the results of the first two steps.

Letting the asterisk denote a proportional change, the system of structural equations is:

$$X^* = aL_X^*, \tag{1'}$$

where a = the elasticity of X output with respect to labor input or labor's imputed share;

$$w_X^* = -gL_X^* + p^*, \tag{2'}$$

where g = the (inverse Marshallian) elasticity of labor demand, equal to capital's imputed share $(1 - a)$ divided by the elasticity of substitution (σ_X);

$$L_X^* = \bar{L}_X^* \tag{3a'}$$

or

$$w_X^* = \bar{w}_X^*; \tag{3b'}$$

$$X^* = -e_1 p^* + e_2 q^* + e_3 \bar{D}^*, \tag{4'}$$

where e_1 = the "own price" elasticity of demand for X, e_2 = the "cross price" elasticity of demand for X, and e_3 = the income elasticity of demand for X;

$$Y^* = bL_Y^*, \tag{5'}$$

where b = analogue of a;

$$w_Y^* = -hL_Y^* + q^*, \tag{6'}$$

where h = analogue of g;

$$L_Y^* = \frac{L}{L_Y} L^* - \frac{L_X}{L_Y} L_X^* - \frac{L_U}{L_Y} L_U^*; \tag{7'}$$

$$Y^* = -e_4 q^* + e_5 p^* + e_6 \bar{D}^*, \tag{8'}$$

where e_4 = analogue of e_1, e_5 = analogue of e_2, and e_6 = analogue of e_3. Equation (8′) is the budget constraint combined with the demand for X function (4). The elasticities (e's) are related as follows: From (4),

$$X^* = -\varepsilon(p^* - q^*) + \eta(\bar{D}^* - q^*),$$

where ε = the "primitive" elasticity of demand for X with respect to relative prices and η = the "primitive" elasticity of demand for X with respect to real income. Then

$$e_1 = \varepsilon, \qquad e_2 = \varepsilon - \eta, \qquad e_3 = \eta,$$

$$e_4 = 1 + \frac{A}{1-A}(\varepsilon - \eta), \qquad e_5 = \frac{A}{1-A}(\varepsilon - 1), \qquad e_6 = \frac{1}{1-A}(1 - \eta A),$$

where $A = Xp/\bar{D}$. Note that $e_1 - e_2 = e_3$ and $e_4 - e_5 = e_6$ and that

$$(e_1e_4 - e_2e_5) = (e_1e_6 + e_3e_5) = (e_2e_6 + e_3e_4) = \frac{\varepsilon - \eta A}{1 - A},$$

$$e_1 = 1 \Leftrightarrow e_5 = 0, \qquad e_2 = 0 \Leftrightarrow e_4 = 1, \qquad e_3 = 1 \Leftrightarrow e_6 = 1.$$

Finally, we have

$$L_U^* = EW(w_X^* - w_Y^*) + EL_X^*, \tag{9'}$$

where E = the elasticity of the unemployment function with respect to its argument and $W = 1/\log(w_X/w_Y)$.

While we have given explicitly neoclassical interpretations to the elasticities, we need not have done so, and the skeptical reader is welcome to furnish his own. For example, one could disbelieve that labor is paid the value of its marginal product and still accept (2) and (6) on the grounds that demand curves for labor exist, and slope downward, but for other reasons. We believe that the model is quite general in that it bears a wide variety of interpretations.

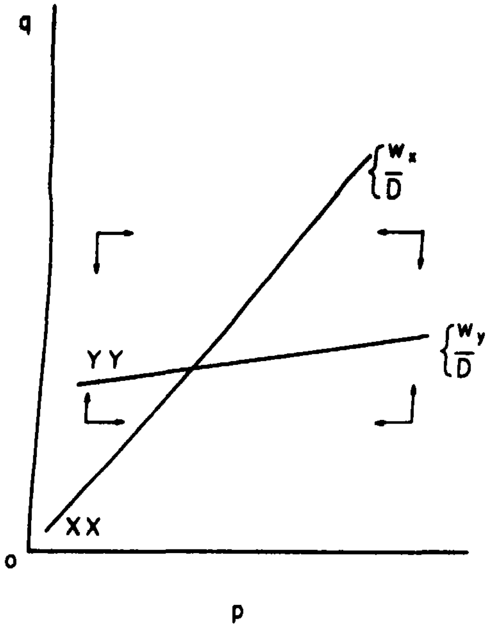

FIG. 2

We first investigate the "product side" of the model. Combining (1′) and (2′) gives a supply of X function with the wage as a parameter. Taken along with the demand for X function, equation (4′), we have, following Hicks, combinations of prices which clear the X market, with the X-sector wage and aggregate demand as parameters (Hicks 1946, chap. 5):

$$p^*\left(\frac{a}{g} + e_1\right) = e_2q^* + \frac{a}{g}\cdot w_X{}^* + e_3\bar{D}^*. \tag{10}$$

Likewise, equations (5′), (6′), and (8′) yield equilibrium price combinations which clear the Y market:

$$q^*\left(\frac{b}{h} + e_4\right) = e_5p^* + \frac{b}{h}\cdot w_Y{}^* + e_6\bar{D}^*. \tag{11}$$

Equations (10) and (11) are depicted in figure 2. The XX curve illustrates (10) and the YY curve (11) on the assumption that XX is more steeply upward sloping. A sufficient condition for this is that e_1e_4 exceed e_2e_5; the product of the "own price" elasticities must exceed the product of the "cross price" elasticities; or, from above,

$$\frac{\varepsilon - \eta A}{1 - A} > 0. \tag{12}$$

We shall regard this as the "normal" case, for if prices rise in response to excess demand, the product markets would be stable. The reader can confirm that if YY is steeper than XX, the equilibrium, if it exists, is unstable.

Solving for p^* and q^* in (10) and (11) gives us

$$p^* = \frac{a(b/h + e_4)}{g\Delta_1} w_X{}^* + \frac{e_2 b}{h\Delta_1} w_Y{}^* + \frac{e_2 e_6 + e_3(b/h + e_4)}{\Delta_1} \bar{D}^*, \qquad (13)$$

$$q^* = \frac{b(a/g + e_1)}{h\Delta_1} w_Y{}^* + \frac{e_5 a}{g\Delta_1} w_X{}^* + \frac{e_3 e_5 + e_6(a/g + e_1)}{\Delta_1} \bar{D}^*, \qquad (14)$$

$$\Delta_1 = e + \frac{e_4 a}{g} + \frac{e_1 b}{h} + \frac{ab}{gh}, \qquad e = (e_1 e_4 - e_2 e_5) > 0.$$

Note that these equations express precisely the consequences of "wage push" ($w_Y{}^*$, $w_Y{}^* > 0$) and "demand pull" ($\bar{D}^* > 0$).

Turning now to the "labor side," we seek combinations of wages which "clear" the respective labor markets. Demand for L_X is given by (1′), (2′), and (4′), with supply of L_X determined by (3a′) or (3b′). Equilibrium combinations of w_X and w_Y are given by

$$w_X{}^* = \frac{e_2}{e_1} q^* + \frac{e_3}{e_1} \bar{D}^* - \frac{e_1 g + a}{e_1} L_X{}^*, \qquad (15)$$

with either $w_X{}^*$ or $L_X{}^*$ given exogenously.

Similarly, for the L_Y market, demand for L_Y is derived from (5′), (6′), and (8′). However, the supply of L_Y must take into account alternate employment in X and unemployment. Taking (7′) and (9′) along with demand, we have

$$\begin{aligned} w_Y{}^* = {} & \frac{e_5}{e_4} \cdot \frac{1}{\Delta_2} p^* + \frac{e_6}{e_4} \cdot \frac{1}{\Delta_2} \bar{D}^* \\ & + \frac{e_4 h + b}{e_4} \cdot \frac{L_U}{L_Y} \cdot \frac{EW}{\Delta_2} \cdot w_X{}^* \\ & + \frac{e_4 h + b}{e_4} \left(\frac{L_X}{L_Y} + E \cdot \frac{L_U}{L_Y}\right) \cdot \frac{1}{\Delta_2} L_X{}^* \qquad (16) \\ & - \frac{e_4 h + b}{e_4} \cdot \frac{\bar{L}}{L_Y} \cdot \frac{1}{\Delta_2} L^*, \end{aligned}$$

$$\Delta_2 = 1 + \frac{e_4 h + b}{e_4} E \cdot \frac{L_U}{L_Y} \cdot W,$$

with either $w_X{}^*$ or $L_X{}^*$ given exogenously.

Equations (15) and (16) are plotted in figure 3. The *xx* curve illustrates (15) and the *yy* curve (16). The *xx* curve is vertical, reflecting the common

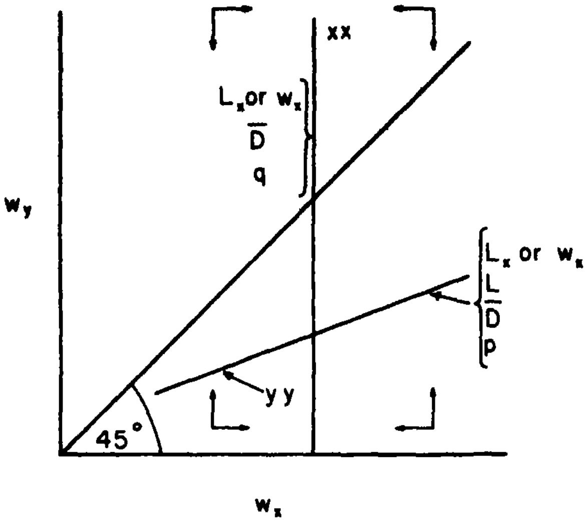

FIG. 3

observation that a wage policy ($w_X^* = \bar{w}_X^*$) or an employment policy ($L_X^* = \bar{L}_X^*$) of labor unions can give rise to equivalent labor market "equilibria." The *yy* slopes upward because higher wages in *X* increase the payoff from being unemployed in search of those wages. Workers in *Y* are thus induced to down tools and become unemployed while searching for employment in *X*. Wages rise in *Y* as workers depart.

Wages are assumed to fall in response to excess supply, and the labor markets are stable. Note, however, that a special interpretation of "disequilibrium" is required in the market for *X* labor. We take either $L_X = \bar{L}_X$ or $w_X = \bar{w}_X$ to be the short-run "equilibrium" supply curves. "Disequilibrium" adjustments are relative to those curves.

The original system of nine structural equations has now been reduced to (13), (14), (15), and (16); unknowns are p^*, q^*, w_X^* or L_X^*, and w_Y^*; exogenous variables are $\bar{D}^*$, $\bar{L}_X^*$ or $\bar{w}_X^*$, and $\bar{L}^*$. Further details of the solution are omitted to save space. The results are set out in table 1. It reports the algebraic sign of the reduced-form (partial) elasticities where the institutional rigidity is a fixed supply, $L_X = \bar{L}_X$. These link proportional changes in exogenous variables with proportional changes in the endogenous ones of chief concern.

As can be seen from the last column of table 1, the comparative static results with respect to $\bar{L}_X^*$ are in most cases ambiguous. An increase in $\bar{L}_X$ ensures that production of *X* will increase and that the wage differential will become smaller. The other results are ambiguous, depending on the sign of some very complicated terms. (Three of these conditions appear in table 1.) The comparative statics with respect to $\bar{D}^*$ and $\bar{L}^*$ are discussed in Section IV below.

TABLE 1

$L_x^{\prime *} = L_x^*$

ENDOGENOUS	EXOGENOUS		
	$\bar{D}^*$	$\bar{L}^*$	L_x^*
w_X^*	+	−	? sign C_1
w_Y^*	+	−	? sign C_2
$w_X^* - w_Y^*$	0	+	−
p^*	+	−	?
q^*	+	−	?
L_X^*	0	0	+
X^*	0	0	+
L_Y^*	0	+	?
Y^*	0	+	?
L_U^*	0	+	? sign C_3

$$C_1 = -\left[ge + \frac{e_4 a}{g} + gh\Delta_1 \cdot E \cdot \frac{L_U}{L_Y} W - e_2 \left(\frac{L_X}{L_Y} + E \cdot \frac{L_U}{L_Y}\right]\right)$$

$$C_2 = \left[(he + e_1 b)\left(\frac{L_X}{L_Y} + E \cdot \frac{L_U}{L_Y}\right) - gh\Delta_1 \cdot E \cdot \frac{L_U}{L_Y} \cdot W - e_5 a\right]$$

$$C_3 = -\left[W\left(he + e_3 b \cdot \frac{L_X}{L_Y} + ge + e_6 a\right) - e\right] \cdot E$$

When we consider the comparative statics when the wage rate is fixed, the results are similar. Almost none of the comparative statics results is determinate with respect to sign. In fact, if we apply Cramér's rule, the determinant of the matrix in the denominator is unsigned. Further, the "economic sense" of conditions which would guarantee a determinate result is not obvious.

The indeterminacy of these results is itself interesting. However, in order to derive results which are more readily amenable to interpretation but which reflect essential features of the model, we impose additional restrictions. We specify that in the neighborhood of equilibrium: (1) the "primitive" elasticity of demand for X with respect to relative prices (ε) is unity; (2) the "primitive" elasticity of demand for X with respect to real income (η) is unity; and (3) the elasticities of substitution in both production sectors (σ_X and σ_Y) are unity. Both demand relations are thus "Graham type" (constant expenditure) functions, and the production functions are "Cobb-Douglas." For convenience, we shall refer to this simple version as the "log-linear model."

The structural equations relating proportional changes are now

$$w_X^* = -(1 - a)L_X^* + p^*, \tag{2''}$$

$$X^* = -p^* + \bar{D}^*, \tag{4''}$$

$$w_Y^* = -(1 - b)L_Y^* + q^*, \tag{6''}$$

$$Y^* = -q^* + \bar{D}^*. \tag{8''}$$

Equations (1′), (3a′) or (3b′), (5′), (7′), and (9′) complete the model.

TABLE 2

	EXOGENOUS			
	$L_X^* = \bar{L}_X^*$	$w_X^* = \bar{w}_X^*$		
ENDOGENOUS	L_X^*	$\bar{D}^*$	$\bar{L}^*$	$\bar{w}_X^*$
w_X^*	−	0	0	+
w_Y^*	sign $-d_1$	+	−	sign $+d_1$
p^*	−	+	0	+
q^*	sign $-d_1$	+	−	sign $+d_1$
L_X^*	+	+	0	−
L_Y^*	sign $+d_1$	sign $+d_1$	+	sign $-d_1$
X^*	+	+	0	−
Y^*	sign $+d_1$	sign $+d_1$	+	sign $-d_1$
L_U^*	sign $+d_2$	sign $+d_2$	+	sign $-d_2$

$$d_1 = U' - 1/\left\{\iota\left[1 - \log\left(\frac{\bar{w}_X}{w_Y}\right)\right]\right\}$$

$$d_2 = \log\left(\frac{\bar{w}_X}{w_Y}\right) - 1 - \frac{L_X}{L_Y}$$

Again, solving market by market, we obtain four equations corresponding to (13), (14), (15), and (16) but which are far simpler and tractable:

$$p^* = aw_X^* + g\bar{D}^*, \tag{13''}$$

$$q^* = bw_Y^* + h\bar{D}^*, \tag{14''}$$

$$w_X^* = -L_X^* + \bar{D}^*, \tag{15''}$$

$$w_Y^* = \frac{\bar{D}^*}{\Delta_2} + \frac{L_U}{L_Y} \cdot \frac{E}{\Delta_2} \cdot W \cdot w_X^* + \frac{1}{\Delta_2}\left(\frac{L_X}{L_Y} + \frac{L_U}{L_Y}E\right) \cdot L_X^* - \frac{\bar{L}}{L_Y} \cdot \frac{1}{\Delta_2}\bar{L}^*, \tag{16''}$$

$$\Delta_2 = 1 + \frac{L_U}{L_Y} \cdot E \cdot W.$$

These structural equations give rise to the reduced-form elasticities. The algebraic signs of these elasticities are recorded in table 2 for future reference.

IV. Applications

A. Demand Management

1) $L_X = \bar{L}_X$. If the institutional rigidity is a restriction of labor supply to the X sector, then, even in the general case, *all real variables are homogeneous of degree zero in aggregate demand,* with prices bearing the whole of the

required adjustment. In this case, even the short-run Phillips curve is vertical (see the first column of table 1).

This is a trivial result in one-sector neo-Keynesian models of the postwar period.[3] If aggregate employment is fixed, so is output, and prices alone absorb demand changes. Simple enough. Not so in our two-sector model, however. Here, the Y sector can respond to demand pressures. Suppose the income elasticity for Y is high relative to that for X. Then an increase in $\bar{D}$ will be channeled mainly into the Y sector, raising the derived demand for Y labor relative to X labor and raising Y wages relative to X wages. The narrowing of the wage differential reduces the profitability of being unemployed, and some of the unemployed give up looking for X-sector employment and enter Y-sector jobs, expanding the output of Y while the output of X remains fixed.

The fact that this cannot occur strikes us as a rather surprising implication of our assumptions. The clue lies in the demand functions which are homogeneous of degree zero in relative prices and real income. The demand relations are satisfied if $\bar{D}$, p, and q rise by equal percentage amounts, with the output of X and Y unchanged. Now, suppose L_X and L_Y are unchanged. Then X and Y are constant, and equal percentage rises in p and q will be reflected in equal percentage rises of w_X and w_Y. With L_X unchanged, there is no reason for L_U, and thus L_Y, to change.

2) $w_X = \bar{w}_X$. We must turn to the log-linear model for interpretable results in this case. See the second column of table 2. An increase in aggregate demand induces an increase in X-sector employment and output, and both prices and the Y-sector wage increase. However, the change in unemployment and the changes in Y-sector employment and output are ambiguous. The change in unemployment depends on the sign of

$$\log\left(\frac{\bar{w}_X}{w_Y}\right) - 1 - \frac{L_X}{L_Y}.$$

Although the sign is ambiguous, we expect it to be negative. The single positive term, $\log(\bar{w}_X/w_Y)$, will be less than 1 if $\bar{w}_X/w_Y$ is less than $e \simeq 2.72$. Then $\bar{w}_X/w_Y$ less than 2.72 is a sufficient condition to ensure that $dL_U{}^*/d\bar{D}^*$ be negative. Accordingly, we assume that unemployment decreases in response to increased demand.

The Y-sector ambiguities turn on the sign of

$$U' - \frac{1}{\{t[1 - \log(\bar{w}_X/w_Y)]\}}.$$

---

[3] For example, see Tobin (1955), who assumes $w_X = \bar{w}_X$. But, according to Tobin, if $L_X = \bar{L}_X$, then the elasticity of aggregate supply is zero.

We interpret this condition below, while noting here that the Y sector will behave as a "countercyclical" sector of "last resort" employment if unemployment is relatively unresponsive to the arguments which determine it: the wage spread and X-sector employment.

To see this, let $\bar{w}_X{}^* = 0$ in equations (13″)–(16″) and let $\bar{D}^* > 0$. Equation (15″) reveals that L_X increases concomitantly, but (16″) tells us that w_Y increases as well. Increases in L_X increase the incentive to become unemployed, while increases in w_Y tend to reduce unemployment by narrowing the wage spread. Unless the wage spread is quite large ($\bar{w}_X$ greater than $2.72 w_Y$), the latter effect dominates the former so that $\bar{D}^* > 0$ implies $L_U{}^* < 0$.

But $\bar{D}^* > 0$ may not have the effect of increasing employment in the Y sector. Recall that $L_Y = \bar{L} - L_X - L_U$. With L_X rising and L_U falling as $\bar{D}$ increases, L_Y can go either way. If the "unemployment effect" ($L_U{}^* < 0$) dominates the "employment effect" ($L_X{}^* > 0$), then L_Y increases. This will occur if the "unemployment effect" is sufficiently large.[4] Unemployment is determined by

$$U\left[tL_X \log\left(\frac{\bar{w}_X}{w_Y}\right)\right].$$

Since the argument of the U function is proportional to t, the magnitude of the unemployment effect is directly related to the exogenous turnover rate and, of course, to the slope of the unemployment function. For a given value of L_X, t determines the number of job openings per period, and hence it directly affects the desirability of unemployment.

If we once again consider the aggregation procedure used to derive the unemployment function, we can say a bit more about the ambiguous comparative statics. In figure 1 the change in J associated with a change from any integer value of L_U to the next is given by the slope of the ray of indifference of the last individual to become unemployed. This implies that the slope of unemployment function at any point is approximated by the slope of the ray of indifference of the last individual to become unemployed. For the jth individual this slope is, from Section IIC above, $C_j = 1/v_j$. Since $\partial v_j/\partial r_j > 0$, $\partial C_j/\partial r_j < 0$. Higher individual rates of discount then decrease the responsiveness of unemployment to J. This then implies that imperfect capital markets tend to contribute to "perverse" results in industry Y because, for any value of J, they reduce the sensitivity of unemployment to changes in J.

---

[4] There is some evidence that perverse or countercyclical employment trends in the Y sector have occurred in both the pre- and postwar periods in the United States. During the Great Depression, the farm sector exhibited some of the characteristics of our underemployed sector. Theodore Shultz observed that in 1932, when urban wages were still considerably higher and falling less rapidly than rural wages, there was a definite reversal of the historical flow of workers from the farm to the city. In fact, 1932 witnessed a net urban-to-rural migration (Schultz 1945). In discussing regulatory problems in the motor-carrier industry during the 1930s, Kahn notes that "with the sharp reductions in industrial employment, large numbers of workers moved into the owner-operation of service stations, farms, trucks, and small grocery stores, so that the supply of labor in these industries actually expanded in the face of declining remuneration" (Kahn 1971, p. 181). Economic conditions have not recently been so severe as to make the identification of occupations of last resort an easy chore. For intuitive support, we can point to the reemergence of street vendors in large numbers; to the countercyclical variations of employment in taxicabs, despite the regulated nature of the industry (Bureau of Labor Statistics, various issues); and to the plight of Vancouver barbers, who complain of continued entry in the face of sharply reduced demand and earnings in the industry.

The slope of the unemployment function will also be quite sensitive to the demographic composition of the labor force. Since $\partial v_j/\partial K_j < 0$, $\partial C_j/\partial K_j > 0$. Assuming that younger people have longer planning horizons, the younger the labor force, the steeper will be the unemployment function for any value of the argument J. We comment on this and related points below.

Before departing our discussion of demand management, a general observation on the effectiveness of demand management in reducing unemployment is in order. The effectiveness of changes in $\bar{D}$ in reducing unemployment are dependent on the slope of the aggregated unemployment function. Other things equal, the greater is U', the more effective is demand management in reducing unemployment. Given the aggregation procedure, the value of U' is dependent on the arrays of individual rates of discount and time horizons. Thus, the effectiveness of changes in $\bar{D}$ as a control on unemployment is sensitive to the demographic composition of the labor force, the degree of imperfection in capital markets, and, as we shall see below, the progress of disadvantaged groups in removing discriminatory barriers to employment. There have been significant changes in at least two of these factors in recent years.

B. Union Power

Referring once again to table 2, an increase in union power ($L_X^* < 0$ or $\bar{w}_X^* > 0$) will decrease employment and output in the X industry. The price of X will rise. Unemployment rises provide, once again, that the wage spread is not too large. But variables which describe the Y-sector response (w_Y, L_Y, Y, q) can move in either direction. Again, the outcome turns on the sign of $U' - 1/\{t[1 - \log(\bar{w}_X/w_Y)]\}$. Only if unemployment is sufficiently insensitive to the arguments which determine it do we predict a rise in Y-sector output and employment and a fall in the Y-sector wage and price.

We focus on the matter of Y-sector employment (L_Y). A rise in X-sector wages (or additional restrictions on labor input) thrusts workers onto the streets ($L_U^* > 0$) or into jobs of last resort ($L_Y^* > 0$). Unemployment rises. If the increase of unemployment is sufficiently small, Y-sector employment will increase. The magnitude of the unemployment response is conditioned by two (possibly) conflicting forces. On the one hand, the reduction of X-sector employment reduces the incentive to become unemployed because fewer new jobs become available for any given turnover rate. On the other hand, the wage differential widens, increasing the incentive to become unemployed. The latter effect dominates so that unemployment increases, but the increase could be small if unemployment is relatively insensitive. If so, then some of the newly unemployed become underemployed in the Y sector.

C. Changing Time Horizons

Substantial demographic changes have occurred and continue to occur in the North American labor force. The impact of these changes in the labor market, and particularly the relationship which these changes bear to the problem of price inflation, has begun to receive attention (Perry 1970; Schultze 1971). Although this is not the place to explore these problems in depth, we can indicate the role that these changes play in our model.

As more women begin seriously to pursue careers, it will pay them to invest in more thorough search activity, that is, in more unemployment in the context of our model. This, of course, implies an upward shift of the aggregate unemployment function. And it also has implications for the structure of unemployment. It implies that for a given level of unemployment a larger proportion of the unemployed will be women, or that the unemployment among women will increase relative to the rate among men.

The relative increase in the unemployment rate of women can be most readily understood in terms of figure 1. The lengthening of women's career horizons would result in more steeply inclined loci of indifference between unemployment and underemployment for women. Recall that $\partial v_j/\partial K_j < 0$, and hence $\partial C_j/\partial K_j > 0$. This implies that at any given level of unemployment the proportion of the unemployed who are women will increase as a result of changing career horizons. If we put this implication slightly differently, the model suggests that a result of increasing aspirations of women is an increase in the unemployment rate of women relative to that of men (see n. 5 below).

The effects of the removal or lessening of racial discrimination are analogous. If "equal opportunity" is interpreted in our model as increasing the probability that members of minority groups can obtain employment in the high-wage sector, then this results in an increased slope of the loci of indifference between unemployment and underemployment for members of minority groups. As with the women's liberation effect, the "equal-opportunity effect" implies an increase in the unemployment rates of minority groups relative to nonminority groups.

In terms of our model, we can think of the postwar baby boom as having essentially two implications. The baby boom obviously represents an increase in the labor force. In tables 1 and 2, the comparative statics are unambiguous. An increase in the labor force has no effect on the X sector and the increase in the labor force is split between increases in unemployment and increases in Y-sector employment. Production of Y increases as a result of the increased supply of labor, and both the wage rate and the price in the Y sector fall.

There is, however, another implication of the baby boom in our model, for it means that the age composition of the labor force has changed. The baby boom means that the proportion of young people in the labor force has increased. Presuming that young people have longer planning horizons, the implication of the increased proportion of young people in the labor force is to increase the slope of the aggregated unemployment function for any given value of the argument J. This is, of course, the same effect which the changing pattern of women's participation in the labor force and the lessening of discriminatory barriers had on the unemployment function. As with the other two effects, the increased proportion of young people in the labor force implies that unemployment among young workers ought to increase relative to the rate among older workers.[5]

V. Conclusions

A feature of our model is that it cannot be used to analyze involuntary unemployment in the usual sense. There is no person in our model economy who cannot find a job *at some wage.* The unemployed have "only themselves to blame," so to speak, in the sense that they have made a rational calculation to improve their economic position. They are *voluntarily* unemployed. Put another way, we have excluded consideration of those persons who do not have, and cannot invent, an occupation of last resort. It is impossible to know whether this group is empirically important, since one cannot measure the bounds of human inventiveness. But if it is true that this excluded group is small, then we must argue that a large measure of observed unemployment is strictly voluntary.

But if unemployment is voluntary in our model, the occupational mix, including job search, is not. Ours might be called a theory of involuntary dislocation. The unemployed (L_U) and the underemployed (L_Y) are prepared to work in the X sector at the going wage, but they cannot readily find employment there. The trouble, of course, lies in the failure of prices and employment to adjust to their competitive levels in the X industry. If they did, we could drop equations (3). Then, if we add the equilibrium condition that $w_X = w_Y$, dislocation (including unemployment) wipes out. Our definition of involuntary dislocation is: *People are involuntarily dislocated if, at current wage rates, they must wait to find a job in their best alternative occupation.*

---

[5] It is perhaps interesting to note that there is evidence which suggests that the implied effect has occurred in all three situations. The unemployment rate among nonwhites relative to whites seems to have increased sharply in the mid-1950s and the ratio of female to male unemployment rates seems to have increased over the decade of the 1960s. Finally, the unemployment rate among young males relative to that among males age 35–44 increased sharply in the late 1960s. The unemployment rates referred to are those from the Bureau of Labor Statistics Household Survey (Bureau of Labor Statistics, various issues).

The consequences for overall economic efficiency of institutional constraints, and resulting involuntary dislocation, are twofold in our model. Resulting price distortions emphasized by Johnson and Mieszkowski and by Diewert occur here as well (Johnson and Mieszkowski 1970; Diewert 1972). But, in addition, there is a dead-weight loss as people become unemployed in an effort to arbitrage the wage gap. Not only are tangency conditions along the production possibilities curve being violated, but equilibrium occurs at an interior point in the production possibilities set which corresponds to full employment of labor.

There is a further loss which is difficult to document but which may be more important: the frustration and sense of injustice felt by those left out of the X industry. In Weizsäcker's pressure group society, these are the citizens who do not know how to organize and/or cannot finance "profitable" pressure groups (Weizsäcker 1972). Those who can, organize; those who can't, protest!

References

Bureau of Labor Statistics. *U.S. Employment and Earnings Statistics.* Washington: Government Printing Office, various issues.

Diewert, E. *The Effects of Unionization on Wages and Employment: A General Equilibrium Analysis.* Ottawa: Dept. Manpower and Immigration, 1972.

Fei, John C. H., and Ranis, Gustav. *Development of the Labour Surplus Economy: Theory and Policy.* Homewood, Ill.: Irwin, 1964.

Harberger, A. C. "On Measuring the Social Opportunity Cost of Labour." *Internat. Labour Rev.* 103 (June 1971): 559–79.

Harris, John R., and Todaro, Michael P. "Migration, Unemployment, and Development: A Two-Sector Analysis." *A.E.R.* 60 (March 1970): 126–43.

Hicks, J. *Value and Capital.* 2d ed. Oxford: Clarendon, 1946.

Holt, C. C. "Job Search, Phillips Wage Relation, and Union Influence: Theory and Evidence." In *Microeconomic Foundations of Employment and Inflation Theory,* edited by E. S. Phelps et al. New York: Norton, 1970.

Johnson, H. G., and Mieszkowski, P. "The Effects of Unionization on the Distribution of Income: A General Equilibrium Approach." *Q.J.E.* 84 (November 1970): 539–61.

Kahn, Alfred E. *The Economics of Regulation.* Vol. 2. New York: Wiley, 1971.

Lewis, W. Arthur. *Theory of Economic Growth.* London: Allen & Unwin, 1955.

Mortensen, D. T. "A Theory of Wage and Employment Dynamics." In *Microeconomic Foundations of Employment and Inflation Theory,* edited by E. S. Phelps et al. New York: Norton, 1970.

Perry, George L. "Changing Labour Markets and Inflation." *Brookings Papers Econ. Activity* 1, no. 3 (1970): 411–41.

Robinson, J. "Disguised Unemployment." *Econ. J.* 46 (June 1936): 225–37.

Rosenstein-Rodan, Paul N. "Problems of Industrialization of Eastern and Southeastern Europe." *Econ. J.* 53 (September 1943): 202–11.

Schultz, T. W. *Agriculture in an Unstable Economy.* New York: McGraw-Hill, 1945.

Schultze, Charles L. "Has the Phillips Curve Shifted? Some Additional Evidence." *Brookings Papers Econ. Activity* 2(1971): 452–67.

Tobin, J. "A Dynamic Aggregative Model." *J.P.E.* 63 (April 1955): 103–15.

Todaro, Michael P. "A Model of Migration and Urban Unemployment in Less Developed Countries." *A.E.R.* 59 (March 1969): 138–48.

Weizsäcker, C. C. von. *The Political Economy of Stability in Western Countries.* Stockholm: Almquist & Wicksell, 1972.

Zarembka, Paul. "Labour Migration and Urban Unemployment; Comment." *A.E.R.* 60 (March 1970): 184–86.

Collective Choice and General Fund Financing

Edgar K. Browning

University of Virginia

This paper develops a geometrical model for the analysis of the collective decision process known as general fund financing. The model is used to compare the outcomes produced under general fund financing and simple majority voting and to explore the comparative statics properties of general fund financing. The major conclusion is that general fund financing often produces unpredictable and paradoxical outcomes. For example, a change in the preferences of voters which would lead to increased outputs of each of two public goods under simple majority voting can produce reduced outputs of both public goods under general fund financing.

It is well known that the equilibrium quantities of publicly supplied goods may depend on the political decision rules governing collective choice. One decision rule, simple majority voting, has been extensively discussed in the literature. However, an equally important decision rule, general fund financing (to be described below), has received relatively little attention.[1] This paper develops a geometrical model which can be used to analyze the comparative statics properties of general fund financing as well as to compare the outcomes it produces with those of simple majority voting. The major conclusion of the analysis is the unpredictability of general fund financing: without an exact knowledge

I would like to thank Roland N. McKean for helpful comments on an earlier version of this paper and Charles J. Goetz for his assistance in straightening out my thinking on the material in Section I.

[1] The most important contributions are Buchanan (1963), Goetz (1968), and Goetz and McKnew (1972). Buchanan's work is the seminal contribution since it was the first to contrast earmarking (simple majority voting) and general fund financing within a public choice framework. The papers by Goetz and Goetz and McKnew, as well as the present paper, also incorporate this important methodological advance.

[*Journal of Political Economy*, 1975, vol. 83, no. 2]
© 1975 by The University of Chicago. All rights reserved.

of voters' preferences there are few a priori predictions that can be made about policy choices under this decision rule.

Building on the earlier work of Buchanan (1963), Goetz (1968), and Goetz and McKnew (1972), the first section of this paper stresses the differences between simple majority voting and general fund financing which are pertinent for a public choice analysis. In the second section the geometry is explained and used to analyze the voting behavior of an individual voter. The application to the group decision process is straightforward and is accomplished in the third section.

I. Simple Majority Voting and General Fund Financing

Since a major portion of the following analysis is devoted to the comparison of general fund financing and simple majority voting, we shall begin with a brief sketch of these decision rules. Actually, the term "simple majority voting" is ambiguous when the output of more than one public good is being determined. It can mean, for example, that the collectivity determines outputs by voting on combinations of the two goods. Thus, a vote might be taken to choose between two missiles plus three police cars and four missiles plus one police car. This is not, however, the usual meaning attached to this decision rule. Normally, simple majority voting refers to determining the equilibrium quantity of each public good by separate votes. The collectivity is viewed as voting on increments to the output of, say, missiles until a majority no longer favors an increase, and then following the same procedure with police cars. It is this second, and more common, meaning of simple majority voting with which we shall be concerned in this paper.

Most simple majority voting models also contain definite assumptions about the relationship between taxes and expenditure levels. It is usually assumed that a certain tax, or set of taxes, is designated to be used to finance whatever level of output is selected. Then a vote to increase the output of a public good is simultaneously a vote to increase taxes in a specified way. It follows that the opportunity cost of more missiles will be viewed by voters as lower disposable incomes (higher taxes) and not as reduced expenditures on other public goods. It can now be seen that the conventional simple majority voting models are all forms of "earmarking" —"the practice of designating or dedicating specific revenues to the financing of specific public services,"[2] together with the provision that the level of the public service be determined by majority voting.

Simple majority voting models are appropriate in cases when taxes are earmarked. These include such important programs as highway expenditures and social security benefits. But a great many, indeed most, expenditures are not financed in this way but are instead dependent on revenues from the "general fund," the size of which is determined independently of the decisions on separate expenditure programs. A different model of collective choice is required for the analysis of this important set of expenditure decisions.

---

[2] Buchanan 1963, pp. 457–58.

It will be simplest to proceed by first describing the model to be used to analyze general fund financing and then explaining why the model captures the essential elements of this decision rule. For analytical purposes, general fund financing is defined as the decision rule whereby the collectivity chooses separately, by majority voting, budget composition and budget size. Choosing the budget composition means determining the proportion of the budget to be devoted to each program. It might be decided that there will be two police cars for every missile, for example. Given this ratio, or "budget mix," a separate decision is made concerning the absolute size of the budget, that is, whether to provide one missile and two police cars or two missiles and four police cars, etc. In effect, general fund financing involves first defining a new public good composed of several goods provided in fixed proportions and then determining the output of this newly defined public good.

It is clear that this procedure does not *describe* the actual process through which expenditure levels are determined by real world governments: legislators do not vote separately to determine the budget ratio. Instead, there are separate votes taken on defense policies, welfare programs, housing programs, and so on. Nonetheless, the model described above is appropriate in most cases that do not involve earmarking. This is because legislators do vote *separately* on taxes, which determine the overall size of the government budget, and on expenditures on individual programs. Separate voting on "expenditure bills" and "tax bills" is the essential characteristic of most government decision making. We will now explain why this procedure produces the same results as the model described above.

Consider a decision to increase expenditures on missiles when tax bills and expenditure bills must be approved separately. The voter will recognize that the government operates within a budget constraint which is determined by the decision on taxes. A vote to increase spending on missiles does not imply that taxes will simultaneously be increased (as in the earmarking case). Instead, the opportunity cost of an increased outlay on missiles will be viewed as occurring at the expense of other items in the budget. Given the government's separately determined budget constraint, a vote to increase spending on missiles is a vote to spend less on other items (police cars). Therefore, although there are separate votes on the items in the budget, the voter will be expressing his preferences on the composition of the budget (the budget mix) since he must recognize that the overall size of the budget is not being determined by these separate expenditure votes; he must recognize the existence of the budget constraint.

When this procedure is used to determine the composition of any given budget, the decision on taxes will determine the overall size of the budget. When voting on taxes takes place, the voter will realize that an increase in taxes need not be spent on particular items he may desire. An increase in taxes will simply enlarge the government's budget constraint, with the determination of the separate items in that budget made in the way outlined above. The voter might approve of a tax increase if it were to be used to finance educational subsidies (earmarking) but oppose it under general fund financing because he expects that the increment in the budget will be allocated to other expenditures he does not approve of. The vote on the tax bill will then be an expression of preference on the size of the budget, given the composition of that budget determined in the separate votes on the many items in the budget.

Thus, the separation of decisions on expenditures and taxes which characterizes most governmental decision processes is equivalent to the separate determination of the budget composition and the budget size. We can show this even more clearly in the geometry developed in the next section.

II. Indifference Contours and Individual Choice

We will assume that the community is composed of three voters, A, B, and C. There are three goods produced, two public goods and one private. The public goods are designated as G and S and the private good as X. It will be assumed that the respective tax shares for each of the two public goods are fixed for each voter.[3] All goods are produced under conditions of constant costs.

The budget constraint of individual A is given by

$$Y^A = P_x X + P_G^A G + P_S^A S, \qquad (1)$$

where Y^A is A's (fixed) money income and X, G, and S are the quantities of the three goods; P_x is the market price (marginal cost) of X, and P_G^A and P_S^A are A's tax prices for G and S. The tax prices will be constant over varying quantities of the public goods since both tax shares and marginal costs are assumed constant, and a person's tax price is equal to the product of his tax share and the marginal cost of the public good.

The budget constraint of individual A is shown as the $T_1 T_2 T_3$ plane in figure 1.[4] Individual A cannot determine the combination of goods on

---

[3] This is a crucial assumption since if tax shares are not fixed no determinate political equilibrium may exist due to the cyclical majority problem. Buchanan (1963) and Goetz and McKnew (1972) also make this assumption, but Goetz (1968) drops this assumption in part of his analysis. Musgrave (1959, pp. 123–24) explains why the simultaneous choice of budget size and tax institution may lead to the cyclical majority phenomenon.

[4] The next several paragraphs discussing the three-dimensional diagram are adapted from my paper (Browning 1974) which uses this construction for the analysis of multiple consumption externalities.

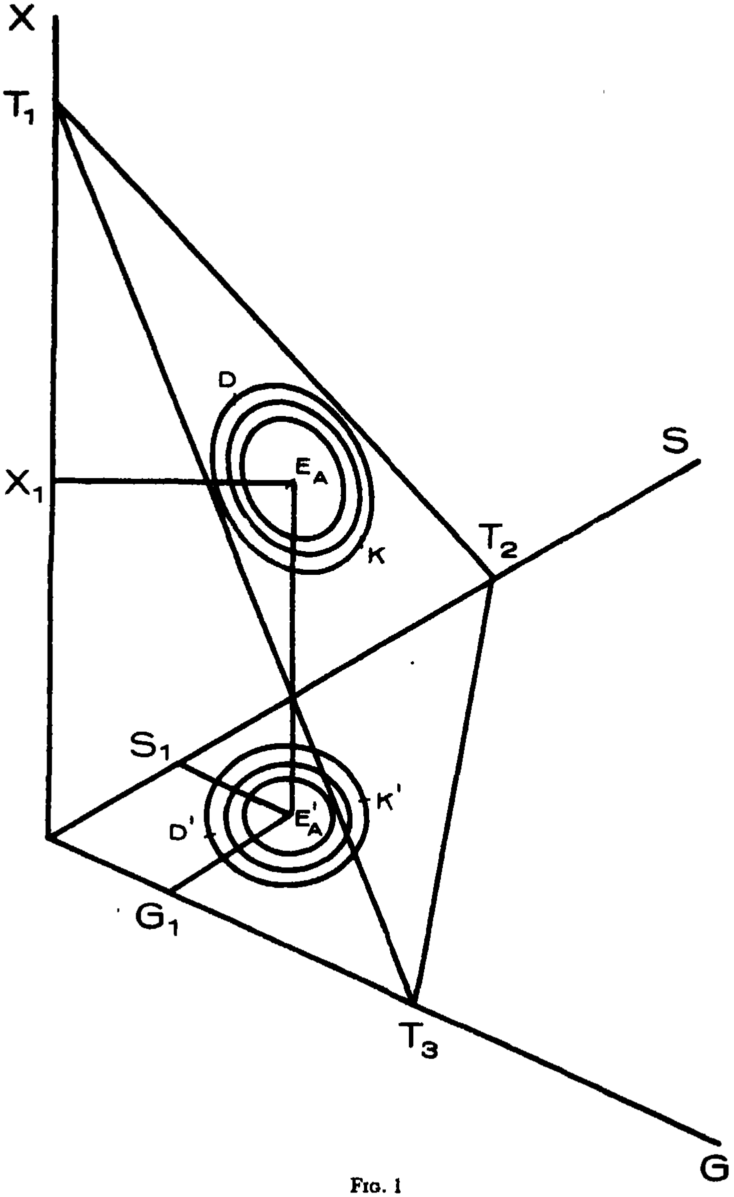

FIG. 1

this plane which he will consume since the quantities of G and S are determined collectively. But since his preferences among these possible combinations of goods will have some impact on the collective determination of G and S via his voting decisions, our first problem is to represent his preferences in figure 1. Since A's utility depends on three variables, G, S, and X, his preference map can be shown as three-dimensional indifference surfaces. No attempt has been made to draw these surfaces in, but they can be visualized as a nest of bowls tilted against the origin in figure 1, with bowls lying further from the origin representing successively higher levels of welfare. (This assumes, of course, that all three goods enter the utility function with positive marginal utilities.) Thus, the highest indifference surface attainable given the budget constraint would appear as a "bowl" just tangent to the budget plane at one point, such as E_A. The allocation of resources at point E_A (X_1, G_1, S_1) is A's most preferred combination of goods possible given the constraint implied by equation (1).

For other combinations of goods on the budget plane the indifference surfaces intersect the budget plane, implying a lower level of utility than the allocation at point E_A. Thus, a bowl lying closer to the origin will have its sides slicing the budget plane in a circular shape (not necessarily a perfect circle), showing all combinations of X, G, and S on the budget plane which imply this lower level of utility. Therefore, lower levels of utility will be shown as concentric closed curves surrounding point E_A on the budget plane.

Individual A's rankings of combinations of goods on the budget plane can be represented by these closed curves. It will be more convenient, however, to show in the two-dimensional G-S space how A ranks the various combinations of public goods attainable given his budget constraint. A two-dimensional treatment contains all the relevant information of the three-dimensional treatment and also permits incorporation of the rankings of several people on the same diagram (as we will see in a moment). Since any point on the budget plane represents a specified combination of G and S, the G-S coordinates of the circular curves can be projected onto the G-S plane. In figure 1, E'_A in the G-S plane shows the most preferred combination of G and S (G_1 and S_1) given by E_A on the budget plane. Similarly, D and K become D' and K' in the G-S plane. The curves in the G-S plane will, of course, retain a circular shape.

Figure 2 reproduces the curves in the G-S plane in more detail. To avoid confusing these curves with indifference curves, they will be called indifference contours.[5] In common with indifference curves, indifference contours reflect the rankings of alternative combinations of goods, but the combinations being ranked are only those lying on a given budget plane. For any given G and S, the budget plane specifies a unique quantity of X so A's preferences can be shown as a function of only G and S as long as the budget plane remains unchanged. (Note that the entire set of indifference contours would shift if the budget plane is changed. A change in the voter's tax prices, for example, alters the

---

[5] Browning (1974) discusses this geometry at greater length and gives a mathematical derivation of these contours.

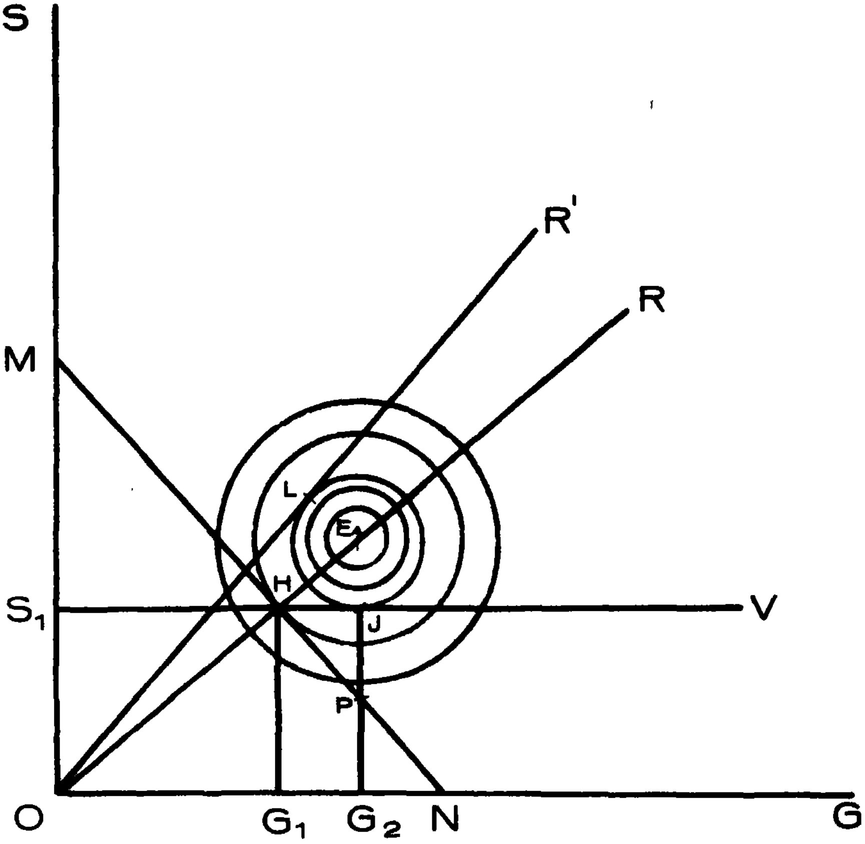

FIG. 2

budget plane and, hence, his indifference contours.) As we move from one point to another on an indifference contour in figure 2, the quantity of *X* is also changing in a specific way (as determined by the budget plane) to keep *A* at the same level of utility.

These concentric indifference contours are the basic geometrical construction which will be used for the analysis of general fund financing and simple majority voting.[6] We can begin by analyzing the voting decisions of individual *A* for several alternative cases. It will then be relatively easy to incorporate individuals B and C into the analysis and examine the group decision-making process.

---

[6] Several other writers have used closed curves such as these in the analysis of voting rules (see Black and Newing 1951; Tullock 1967; and Buchanan 1968). None of these authors, however, explains the derivation of these curves (which I believe is very helpful in understanding what determines their shape and position), nor do they use them for the analysis of general fund financing. However, Buchanan does devote a brief section in his textbook (1970, pp. 131–34) to general fund financing, but without using this terminology.

Under simple majority voting, individuals vote on the quantity of one public good at a time. Suppose that the vote is on the quantity of G: given some quantity of S, say S_1, individual A must rank all of the alternative quantities of G. The alternatives permissible under this voting rule are shown by the horizontal line S_1V. Individual A's most preferred combination of S and G along S_1V is shown by point J where an indifference contour is tangent to S_1V. Given any pair of alternatives, both of which involve smaller quantities of G than G_2, A would prefer (vote for) the larger, and for alternatives involving larger quantities of G he would prefer the smaller. His most preferred quantity of G may depend on, of course, the quantity of S being provided. For another quantity of S, a different horizontal constraint is relevant for A's voting decision, but the central point is that under simple majority voting the voting determination of the quantity of G always involves choices lying along some horizontal constraint. Similarly, the voting determination of the quantity of S involves choices lying along some vertical line. For a given quantity of G, A's preferred quantity of S occurs at the tangency between an indifference contour and the relevant vertical line.

Now let us examine A's voting behavior under general fund financing. Suppose the taxing decision has been made and the revenues expected will finance the combinations of G and S shown by MN: this is the government's budget constraint, with a slope equal to (minus) the ratio of the marginal costs of the two public goods. Given the alternatives along MN, individual A prefers point H with quantities G_1 and S_1, that is, point H is his preferred budget mix for the budget size given by MN. Starting at point H, suppose a proposal is made to increase the quantity of G from G_1 to G_2. With general fund financing as the decision rule, the individual will vote against this proposal because approval of this proposal does not imply that the allocation at point J will be implemented. To achieve the allocation at point J requires a larger total budget, and the budget size is independently set by the decision on the tax bill. Instead, increasing the output of G to G_2 within the separately chosen budget constraint implies moving from H to P: the increase in G must come at the expense of S. Thus, he would vote against the proposal to increase the output of G since, in the absence of earmarking provisions, this expenditure decision is not tied to a decision to increase taxes.

When voting on the output levels of G and S under general fund financing, then, the voter is expressing his preferences on the budget mix. Point H is the preferred budget mix for the budget size given by MN. The ray OR shows his preferred budget mixes for alternative sizes of the budget and is derived by connecting the tangencies between the indifference contours and budget constraints for alternative sizes of the budget. For convenience, this locus of preferred budget mixes will be drawn as a straight line implying that the preferred ratio of G to S is independent of the absolute size of the budget. This assumption is made for convenience of exposition and can be dropped without altering any of our conclusions.

We have seen how the individual's vote given the size of the budget will reflect his preferences concerning the composition of the budget. The other aspect of general fund financing is the decision on the size of the budget itself. For purposes of this decision, the budget mix is taken as given. Choosing the size of the budget involves choosing a point on the ray OR, if this is the budget mix selected by the community. Individual A would then prefer the budget size indicated by point E_A. Of course, confronted with a different budget mix his preferred point would, of necessity, change. If the budget mix is shown by OR', he would prefer (vote for) the combination at L where an indifference contour is tangent to OR'.

Before proceeding with the group case, figure 2 can be used to illustrate some of Buchanan's conclusions.[7] Buchanan compares earmarking (what we have called simple majority voting) and general fund financing for the case when one voter possesses the median preferences in the community for each public good taken separately as well as for the budget size under general fund financing. With simple majority voting the collective outcome will then be A's individual optimum, E_A. Buchanan then proceeds to show that the outcome under general fund financing will normally be inferior to this. Assuming that an external "budgetary authority" sets the budget mix, he shows how the outcome will be affected. In figure 2 if the mix is set as OR', and if individual A is the median voter on budget size, the equilibrium is at point L. The combination of public goods at point L is inferior to the equilibrium bundle under simple majority voting, at least according to A's preferences. (Buchanan also derives elasticity conditions under which the budget size is larger under general fund financing.)

As Goetz (1968) has emphasized, a major defect in Buchanan's analysis is the assumption that the budgetary mix is exogeneously determined. There seems no compelling reason why the budget mix would be immune to the political forces which determine the other variables. Thus, a logical extension is to have the budget mix, as well as the other variables, determined by majority voting.[8] This is easily analyzed by using the geometrical construction of indifference contours.

---

[7] See Buchanan 1963.

[8] Goetz and McKnew (1972) have made this extension by giving a specific numerical example based on the determination by majority voting of both budget mix and budget size. The model developed in the present paper can be used to demonstrate their results in a somewhat simpler fashion.

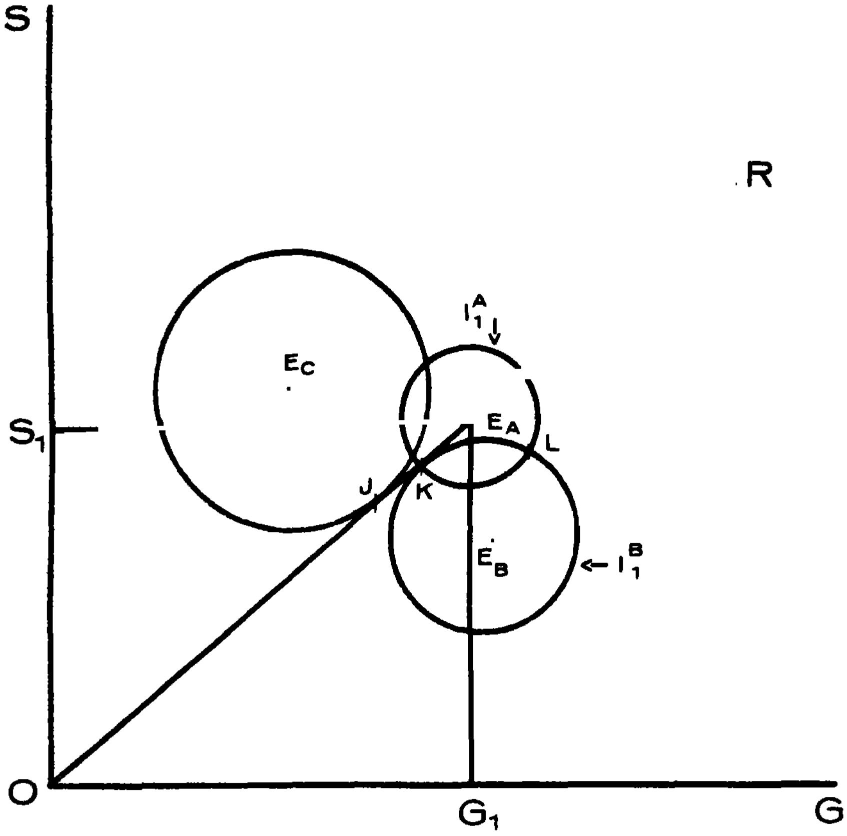

FIG. 3

III. Group Choice

In figure 3 the individual optima of the three voters are shown as E_A, E_B, and E_C. When simple majority voting is used to determine the quantities of the public goods the equilibrium combination is E_A since, given the configuration of preferences assumed, individual *A* is the median voter on both issues.[9] Now consider how the equilibrium would be affected if general fund financing were used as the decision rule. Individual *A* would still control the selection of the budget mix since the ray *OR* passing through E_A is the median preferred budget mix. Given the alternative budgets along *OR*, the preferred budgets of *C*, *B*, and *A* are, respectively, J, K, and E_A. Therefore K, the median preferred budget, will be selected. Note that the quantities of *both* public goods are smaller under general fund financing than under simple majority voting. Another point to note is that while individual A is the median voter on the budget mix individual B is the median voter on budget size. Many of the peculiarities of general fund financing result from the fact that different voters may cast the controlling vote on the two issues.

---

[9] This is necessarily the equilibrium only if the indifference contours are perfect circles. (Note that if *C*'s contours are ellipses one of them could be tangent to the vertical line dropped from E_A, in which case *C* would be the median voter on the quantity of *S*.) All of the relevant indifference contours to show the equilibrium under simple majority voting have not been drawn in to avoid cluttering the diagram unnecessarily. For a lengthier discussion of simple majority voting using this type of geometry, see Buchanan (1968), chap. 6.

The analysis of figure 3 illustrates how the use of indifference contours permits us to identify and compare the equilibria under simple majority voting and general fund financing. But the specific conclusion that the quantities of both public goods fall under general fund financing is not a necessary consequence of that decision rule. On the contrary, it is equally possible that the quantities of both public goods will be greater under general fund financing, or they may be identical, or the quantity of either good may increase and the other fall. The exact consequences of switching to general fund financing depend not only on the decision rule but also on the configuration of indifference contours. (The reader may wish to experiment with different configurations of the curves to verify the possibilities mentioned above.) Since there is no way to rule out a priori any particular configuration of indifference contours, it appears that it is impossible to make any definite prediction about the effects on the quantities of the public goods of switching from simple majority voting to general fund financing. Only with knowledge about the exact pattern of indifference contours can definite predictions be made.

For point K to represent a stable equilibrium under general fund financing it must be assumed that voters do not behave strategically and do not engage in logrolling. Unfortunately, this assumption may not be fully appropriate. To see why a voter may have incentive to behave strategically, consider individual A who casts the pivotal vote on the budget mix. It was assumed that he would favor the mix indicated by OR which passes through his optimum, E_A. However, by confronting individual B with a different budget mix, A can affect the combination of goods B will vote for, and hence affect the final outcome since B is the median voter on the budget size. For instance, suppose A votes for a budget mix shown by a flatter ray from the origin which passes through the lens-shaped area formed by the intersection of I_1^A and I_1^B at points K and L. If B prefers a combination of G and S lying within the lens-shaped area when confronted with this ray, then A will be better off than if he had chosen OR and ended up at point K. Thus, voters may have incentive to use strategy in their voting patterns under general fund financing.

There is also nothing in the institution of general fund financing which prevents logrolling. For instance, A might offer to B to vote for a budget mix lying within the lens-shaped area in return for B's agreement to support a budget size falling within that area. Since both A and B are better off with combinations of G and S in that lens-shaped area, there is incentive for such exchanges to occur. Logrolling can also take place between A and C or between C and B. It is not clear what type of equilibrium, if any, would be established.

When there is a large number of voters, and in the absence of collusive arrangements, strategic behavior and logrolling are unlikely to occur because of the imperceptible impact any one person's vote has on the final outcome. For that reason we will proceed under the assumption that these phenomena do not occur. But the possibility of strategic behavior and logrolling should not be completely dismissed: most policy decision making takes place in legislatures, and the number of voters there is probably small enough to permit such behavior. (Of course, these phenomena can also arise under simple majority voting.)

These examples are sufficient to indicate the difficulty of comparing general fund financing and simple majority voting with respect to the quantities of the public goods provided. But another way the model can be used is to generate predictions of the effects of a change in the underlying parameters given general fund financing as the decision rule. For instance, a change in tax shares, incomes, or resource supplies (to mention only the obvious possibilities) will alter the budget plane in figure 1 and hence modify the indifference contours. Our problem is to determine what consequences such changes will have on the budget.

One example will be given which will show, we believe, how unpredictable are the comparative statics properties of general fund financing. Assume that the preferences of all three voters change so that the optimum quantity of G for each voter is increased. We will show, however, that the end result may be a reduced equilibrium quantity of G. In figure 4 the individual optima for the initial sets of preferences are shown as E_A, E_B, and E_C. The budget mix is given by OR and the equilibrium is at E_A since A is the median voter on the size of the budget; G_1 and S_1 are the initial quantities of the public goods.

When the preferences of the voters change, the individual optima shift to E'_A, E'_B, and E'_C, each implying a larger quantity of G than before. Given these new indifference contours the new budget mix is shown by the ray OR' since B is the median voter on the budget mix. (As drawn, the change in the budget mix favors good G, but it is possible for the new budget mix to involve less G per unit of S.) The preferred budget sizes are J, K, and E'_B, so A is the median voter and K is the equilibrium budget. The quantities of both public goods are smaller at K than at E_A, despite the fact that all three voters desired a larger quantity of G. Thus, if we know only that the demands for G by all three voters have increased, we cannot predict that the equilibrium quantity of G will necessarily increase (although it may increase, of course, with a different configuration of indifference contours).

We assumed that the change in the indifference contours was brought

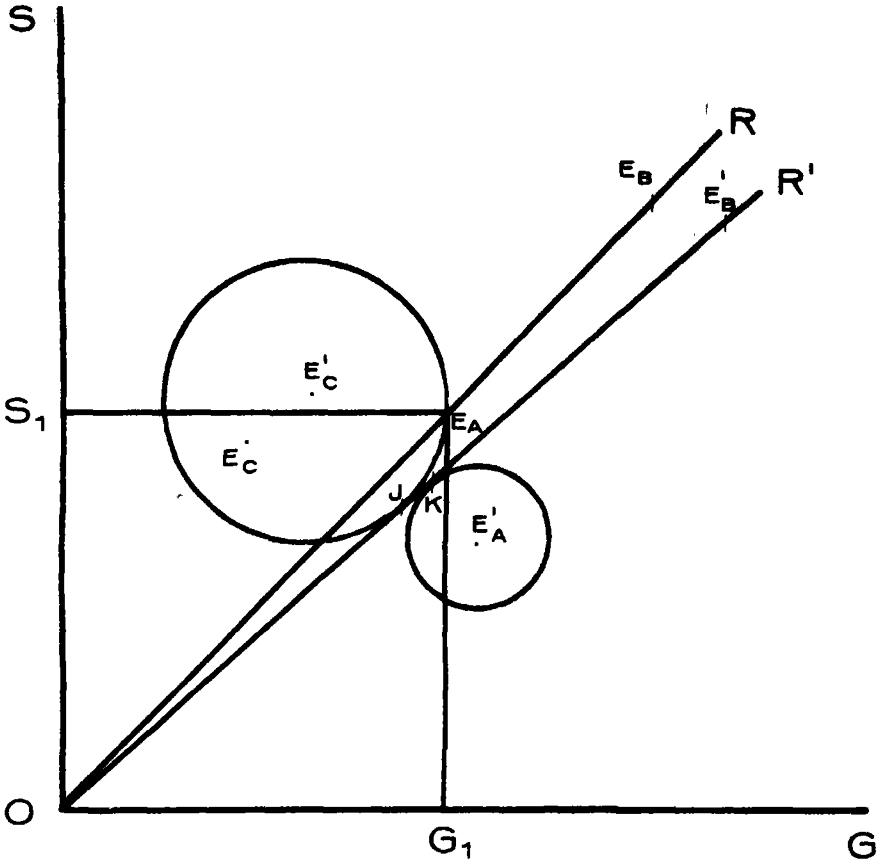

Fig. 4

about by a change in voters' preferences. Other changes in the underlying parameters could also produce the same type of shifts in the indifference contours. For example, a reduction in the marginal cost of producing *G* would shift the budget plane in figure 1 in a manner consistent with the shift in the indifference contours shown in figure 4. (At a lower tax price the optimum quantity of *G* for each voter would be greater, but the optimum quantity of *S* could go up or down, depending on whether it is a complement or substitute for *G*.) Also, a matching grant for the provision of *G* from a higher level of government could also shift the budget planes in this way.[10] Thus, it is also impossible to make an a prior prediction about the effects of these changes in the underlying parameters.

One other interesting point is illustrated by figure 4. If simple majority voting were the decision rule, the initial equilibrium would be E_A, the same as with general fund financing. But following the shifts in the indifference contours, the quantities of both public goods would increase under simple majority voting since A would be the median voter on G (and he prefers a larger quantity than G_1) and C would be the median voter on S (preferring a quantity larger than S_1).[11] Therefore, a shift in the indifference contours which would increase the quantities of both public goods under simple majority voting can reduce the quantities of both public goods under general fund financing.

---

[10] This is the example discussed by Goetz and McKnew (1972). Thus, fig. 4 can also illustrate their result that a matching grant to a local government may be less stimulative than a lump sum grant.

The major theme of this analysis has been the difficulty of deriving predictions about the working of general fund financing. Apparently, almost any outcome is possible. This is a disturbing conclusion for analysts of the political process, especially in view of the claim by Goetz and McKnew (1972, pp. 225–26) that their model (which is based on the same assumptions as the one used here) "exactly reflects the customary government practice of voting separately on taxes and expenditures under so-called general funds financing." If this is correct, then we are a long way from a thorough understanding of how government policy is affected by such factors as changes in preferences, the decision rule, technology, or resource supplies.

References

Black, Duncan, and Newing, R. A. *Committee Decisions with Complementary Valuation.* London: William Hodge, 1951.

Browning, Edgar K. "The Diagrammatic Analysis of Multiple Consumption Externalities." *A.E.R.* 64 (September 1974): 707–14.

Buchanan, James M. "The Economies of Earmarked Taxes." *J.P.E.* 71, no. 5 (October 1963): 457–69.

———. *The Demand and Supply of Public Goods.* Chicago: Rand-McNally, 1968.

———. *The Public Finances.* Homewood, Ill.: Irwin, 1970.

Goetz, Charles J. "Earmarked Taxes and Majority Rule Budgetary Processes." *A.E.R.* 58 (March 1968): 128–36.

Goetz, Charles J., and McKnew, Charles R., Jr. "Paradoxical Results in a Public Choice Model of Alternative Government Grant Forms." In *Theory of Public Choice,* edited by J. M. Buchanan and R. D. Tollison. Ann Arbor: Univ. Michigan Press, 1972.

Musgrave, Richard A. *The Theory of Public Finance.* New York: McGraw-Hill, 1959.

Tullock, Gordon. "The General Irrelevance of the General Impossibility Theorem." *Q.J.E.* 81 (May 1967): 256–70.

---

[11] Once again, this assumes circular indifference contours (see n. 9 above).

Communications

Keynes and the Economic Activists of Pre-Hitler Germany

George Garvy

Federal Reserve Bank of New York

One of the perplexing riddles in the history of social science is how a man of the intellect of Keynes could have labored for years on what he considered to be a revelation without becoming aware of its multifarious antecedents, and how such a large segment of the English-speaking community of economists could have accepted his analysis and policy conclusions as such.

Actually, two separate questions are involved. One is why Keynes ignored the numerous antecedents for his theory of determination of aggregate demand, which became associated with his name. The other is why he lacked interest in countercyclical policy proposals advanced by economists of the two countries most severely affected by the Great Depression.

This note focuses on economic policy guides which have entered history as "Keynesian," rather than the theoretical underpinning for these prescriptions, which he gave in *The General Theory of Employment, Interest, and Money* (hereafter referred to as *General Theory*; first published in 1936, reprinted in Keynes [1973]). Neither is it my purpose to trace the antecedents of *General Theory* in the history of economic thought, or the interpretation of either theory or policy conclusions by the followers of Keynes.

The author gratefully acknowledges comments and suggestions on an earlier draft made by Professors A. P. Lerner, A. Lowe, and J. Marschak and Dr. Walter Salant. The author is responsible for all possible errors of fact and interpretation, and for the remaining irreverence. All the letters quoted are in the author's collection. The views expressed are solely those of the author and not of the Federal Reserve Bank of New York.

[*Journal of Political Economy*, 1975, vol. 83, no. 2]

© 1975 by The University of Chicago. All rights reserved.

A few words on the first question will, however, help to put the second in proper perspective.

The few random references to forerunners can be found in the anticlimactic chapter "Notes on Mercantilism, the Usurer Laws, Stamped Money and Theories of Underconsumption" preceding the concluding chapter of *General Theory*. Accusing academic economists, including himself, of ignoring the "strange, unduly neglected prophet Silvio Gesell" (Keynes 1973, 7:353),[1] Keynes devoted a substantial section in chapter 23 to him, concluding that "the idea behind stamped money [one of the two main policies advocated by Gesell] is sound" (7:357). He added that "the future will learn more from the spirit of Gesell than from that of Marx" (7:355).

One wonders why Keynes devoted five pages to the life and theories of a typical monetary crank, whose name has in the meantime lapsed into oblivion,[2] while ignoring the numerous economists with "heretic" views in the very area in which he believed himself to be breaking a new path. Comments in the two volumes of *Collected Writings* (1973) devoted to *General Theory* provide no clue, and Gesell is not even mentioned in the index to them. The only answer I am able to suggest is that Gesell, who presented a fairly complete theoretical structure, was one among the "four heretics" who were the subject of H. T. N. Gaitskell's essay (1933) in Cole's book, published when Keynes was working on *General Theory*. Except for Gesell, the only German, French, or Russian writers mentioned in *General Theory* are from two secondary sources—Heckscher and Mitchell—but none is discussed at any length.

In more than one instance, Keynes failed to pursue leads which would have taken him to sources of what he later believed to be discoveries of his fertile mind. Thus, in his well-known reply to Hayek ("The Pure Theory of Money," published in 1931, reprinted in Keynes [1973, 13:243 ff.]), Keynes conceded that the neglect of Wicksell by English prewar economists was "mistaken" (but does this exonerate the post–World War I Keynes?). Yet, he managed to give, in early autumn 1936, shortly after the publication of *General Theory*, a lecture at the University of Stockholm in which, judging from the published version of the same year ("The Theory of the Rate of Interest" [1973, 14:101 ff.]), he did not even mention Wicksell, even though a translation of his classic from the German by R. F. Kahn had just been published.[3]

---

[1] While best known in Germany, Gesell was born and lived most of his life elsewhere. In the first chapter (Keynes 1973, 7:32) Gesell is bracketed with Karl Marx and Major Douglas as part of the economic "underworld." To this, Schumpeter (1936) reacted forcefully: "I am no Marxian. Yet I sufficiently recognize the greatness of Marx to be offended at seeing him classed with Silvio Gesell and Major Douglas." Schumpeter passed away before the author of *An Essay on Marxian Economics* revealed that Keynes "never managed to read Marx" (Robinson 1973, p. ix). In the *Treatise* (1930, 2:215–33), but without mentioning any names, Keynes had already made reference to the views of the "army of heretics and cranks" who advanced monetary panaceas that boiled down to (inflationary) credit creation, while "a principal object" of the *Treatise* was to define "the time criterion of credit which was to be non-inflationary." See also Dillard (1942).

[2] While the pre–World War II *Encyclopedia of the Social Sciences* carried a short article about Gesell, its successor, the *International Encyclopedia of Social Sciences* (1968), dropped the entry.

A simple answer to the question of why *General Theory* was written in an intellectual vacuum can be given by referring to the well-known isolation of Cambridge economists within the proverbial insularity of Great Britain. Essentially, Keynes fed on Keynes; this has been made abundantly clear to the uninitiated by the recent publication of two volumes of his *Collected Writings* (1973) devoted to *General Theory*. In the years when it was being written, his peers in Cambridge and Oxford, and a small, select group of faculty and students, meeting as the Cambridge Economic Club at his rooms at King's College, were both the testing ground for his ideas and the opportunity for limited feedbacks from the outside world.

Keynes read little of what contemporary economists had to say (unless they submitted manuscripts to the *Economic Journal*, of which he was the editor). Nor did he dig deeply into the wealth of ideas and analyses of the preceding generations of economists. He largely ignored the contributions of his contemporaries (and the generations preceding them) who published in languages other than English.[4] What a contrast to the author of the other historically significant major endeavor to present a general theory of capitalism—Karl Marx.

---

[3] In the *Treatise* (1930 1:199), Keynes made reference to the school of German and Austrian writers "whose theory on bank-rate in relation to the equilibrium of savings and investment, and the importance of the latter to the credit cycle, is fairly close to the theory of this *Treatise*." Conceding (in a footnote) that he "should have made more reference to the work of these writers" (Hans Neisser being one of those specifically mentioned), Keynes gives as reasons for ignoring them the fact that his attention had been drawn to these writings only after the *Treatise* was already in the hands of the printer, and also his poor knowledge of German. A natural scientist aware of the existence of relevant sources could hardly use such an excuse and keep the respect of his peers. Keynes did not investigate the matter further in the *General Theory*, on which he began to work shortly after the publication of the *Treatise*. When Neisser visited Keynes in the fall of 1932, he apparently did not take the opportunity to acquaint him with the views of German advocates of deficit spending. In the Keynes obituary article (Neisser 1946), when pointing out that deficit spending as a road to full employment had been advocated in various quarters before the appearance of *General Theory*, Neisser referred only in general terms to the contributions of the "non-Austrian branch of the 'neo-Wicksellian School'" without specifically referring to his own role and the efforts of like-minded German economists.

[4] Keynes's ignorance of relevant literature, some of which was in foreign languages, is usually cavalierly dismissed by his admirers. Seymour Harris's (1955) comment is typical: "Keynes probably did not read as much of the economics written as he should have. He was too busy [laboriously distilling what others had formulated before him?] to do so. Keynes paid inadequate attention to these developments though in his *Treatise* he showed that he was aware of some of Wicksell's contributions." Not one of the 40 articles in Harris (1947), devoted to the examination of Keynes's contribution to economics, examines or even mentions the limitations of his intellectual horizon. Suprisingly, Schumpeter, a man of encyclopedic knowledge, made no reference in his contribution

By and large, those American academics who had become apostles of the "new economics" preferred to bask in the glow of the "Keynesian revolution" rather than to filter the light coming from Cambridge through the prism of analysis already embodied in the writing of their own colleagues. It remained for Alvin Hansen to make a major contribution toward putting *General Theory* in the proper historical perspective.[5]

The continuing limitations of the Cambridge horizon are illustrated by the monumental and erudite edition of Keynes's writings (1973) undertaken by his close associates and sponsored by the Royal Economic Society. It has failed completely (judging from the volumes so far published, including those devoted to *General Theory*) to relate Keynes's contribution to those of his predecessors or his contemporaries who had developed similar theoretical concepts, analytical approaches, or policy conclusions.[6] However, what matters for the forward march of mankind is not who deserves credit for inventing any particular analytical apparatus or set of policy prescriptions, but who succeeded in making them widely accepted. This is the lasting contribution and impact of Keynes, who changed the entire focus of economic analysis.

Ever since the publication of *Economic Consequences of the Peace* (1919), Keynes was anxious to influence economic policy.[7] Long before providing a generalized theoretical basis for compensatory fiscal policy, he had been advocating public works. More than many of his contemporaries in the economic profession, he was aware of the similarities in the problems

---

to the parallelism between Keynes's views and those developed in other countries, and in particular to policy discussions in Germany to which he was exposed just before emigrating to the United States. In his posthumous work (1954), Schumpeter noted that "in the 1930's other works appeared that, each in its own way attempted to express views that were similar to Keynes's in important points." In this short preliminary and incomplete sketch, included in the last chapter, the *General Theory* is characterized as "a feat of leadership" and "the greatest literary success in our epoch," while its lack of originality is more than hinted at in several passages.

[5] Hansen, who already in 1927 had provided a wide-ranging survey of European contributions to business cycle theories, including a chapter on monetary theories with a wealth of references to Keynes's predecessors, later (1951) expanded some of this discussion, with particular emphasis on the theories of the investment demand function. In 1952, Hansen wrote that with respect to the general problem of aggregate demand and the factors that determine the level of employment, "Keynesian theory is certainly not altogether new."

[6] In addition to Hansen, several other non-British economists, including Grotkopp (1954, esp. pp. 232 ff.), long ago drew attention to the most important—and some lesser known—forerunners of Keynes. More recently, Gunnar Myrdal (1973) wrote: "In Sweden, where we grew up in the tradition of Knut Wicksell, Keynes's works were read as interesting and important contributions along the familiar line of thought, but not in any sense as a revolutionary breakthrough."

[7] The extent to which Keynes actually influenced the policies of President Roosevelt is a matter of dispute. Thus, Marriner Eccles (who in June 1934 became chairman of the Board of Governors of the Federal Reserve System) commented as follows on a strategy meeting held prior to Roosevelt's inauguration: "With the exception of Ezekiel and Tugwell, I doubt whether any of the men in the room had ever heard of John Maynard Keynes, the English economist who has frequently been referred to as the economic

which advanced industrial countries faced with the onslaught of what we now call the Great Depression. Why did he show so little interest in the ideas and programs put forward in the early thirties by important groups of economists in the two countries in which countercyclical policy was the main domestic policy issue?

In the years between the crash of 1929 and the appearance of *General Theory*, a large and distinguished group of American economists advocated policies designed to lead the country out of the depression. Similar efforts were made in several countries in Europe, including Germany. True, like much of what Keynes advocated in numerous articles and letters to the press during the twenties and early thirties, the action proposals of his American and German colleagues were calls for action. These policy proposals were designed to deal with a given situation, and their authors typically argued that such policies were both desirable and feasible. Generally, they were not presented within a framework of an elaborate and original theoretical analysis, even though in some cases advocacy of expansionary policies was rooted in theoretical analysis advanced earlier by their respective authors. And, indeed, there were considerable differences in the reasoning behind individual anticyclical policy proposals.[8]

Most of those who advanced "Keynesian" ideas in the United States were well established and in the mainstream of academic thought. Why the income-expenditure debate among American economists had not reached the venerable halls of King's College is a problem which clearly deserves more research, since language was not a barrier. J. Ronnie Davis's prize-winning monograph (1971) and the earlier study by Herbert Stein (1969)[9] make it abundantly clear that Keynes's policy proposals

---

philosopher of the New Deal. At least none of them cited his writing [*sic*] to support his own case, and the concepts I formulated, which have been called 'Keynesian,' were not abstracted from his books, which I had never read" (1951, pp. 131–32). Rexford G. Tugwell, a member of the Brain Trust, also vigorously denied any influence of Keynes in shaping Roosevelt's economic policies. Tugwell claims to have arrived at the fundamental economic policy approaches of the New Deal "before he ever heard of Keynes." For a fuller discussion, see Tugwell (1969, pp. 373–75).

[8] It should be noted at this point that in Germany, Say's Law did not occupy the same undisputed position as it did with Anglo-Saxon economists, and therefore less need was felt to demolish this central proposition of classical economics to justify compensatory fiscal policy.

[9] Trescott agrees that the income-expenditure theory, which became textbook Keynesian analysis, was "pretty well formed by Americans in the 1930s" (1973, p. 728). Morris A. Copeland, who was one of the first and perhaps the first American economist to challenge the validity of Say's Law (Copeland 1931), feels that Davis's statement goes too far and should be modified as follows: "By 1935, most of the leading pre-Keynesian economists [in the United States] timidly embraced the objective of compensatory public finance and timidly endorsed the policy as well as its reasoning." He suggests that the reason for the timidity was "that there was no accepted theoretical base for a compensatory fiscal policy" (personal communication, June 30, 1974). Ayres commented in 1946: "As everybody knows, the movement which is now commonly identified with his [Keynes's] name was in full swing before the appearance of *General Theory*." See also Sweezy (1972).

cannot be regarded as innovative when seen in the context of the earlier discussions in the United States.

When *General Theory* appeared, the view of American economists, as Gordon Tullock pointed out in his foreword to Davis's book, was already "overwhelmingly Keynesian."[10] In the words of Davis: "By 1935 most of the leading pre-Keynesian economists [in the United States] solidly embraced the objective of compensatory public finance and firmly endorsed the policy as well as its reasoning," while "as late as mid-1934 Keynes was still a pump-primer" (1971, p. 144; see also pp. 37, 141).[11]

Keynes clearly had not taken advantage of various opportunities for getting acquainted with the views of the U.S. proponents of an active policy to overcome the Great Depression. This is particularly surprising in view of his attendance at the Harris Foundation lectures at the University of Chicago in June 1931, at which he gave two lectures on "An Economic Analysis of Unemployment" (1973, 13:343 ff.). He also participated in several conference seminars, including one in which the question "Is it possible for governments and central banks to do anything or propose to remedy unemployment?" was debated and during which he advanced ideas that later became the cutting edge of *General Theory*.[12] One would vainly look in the index to volumes 13 and 14 of Keynes (1973) for a reference to any of the American economists whose views are discussed by Davis (1971) and Stein (1969).

Keynes also showed no interest in the discussion among German economists on the need for countercyclical monetary and fiscal policies designed to increase aggregate demand, much of which preceded *General Theory* by 5 years.

The countercyclical policies advocated by various individuals and groups who in Germany had, to various degrees, advocated "Keynesian" policies in the early thirties, the theoretical framework on which they were based, as well as their publications and activities to mobilize public opinion in support of their views, still await the historian.[13] So does elucidation of the question of to what extent precisely they anticipated what has become known as "Keynesian economics," with its stress on countercyclical fiscal policy.[14]

---

[10] However, as Trescott (1973) points out in his review of Davis's book, Roosevelt did not get his advice from these economists, nor did he even consider fiscal policy a recovery measure until he came under the influence of L. Currie and M. Eccles.

[11] Several lists of economists who had endorsed policy recommendations along these lines, given in Davis (1971, appendices A, B), read like a *Who's Who* of American economists. Keynes seems to have been better acquainted with U.S. economists of a conservative persuasion. Thus, in late summer 1928, he sent draft copies of his article "Is There Inflation in the United States?" to the following: C. H. Bullock, W. R. Burgess, W. M. Parsons, C. Snyder, and B. Strong (Keynes 1973, 13:52 *n*.).

[12] The only participant from Germany, Otto Nathan, at that time a high official in the Statistical Office, was then an active opponent of programs to foster employment by expansionary fiscal policy.

[13] Grotkopp wrote wistfully: "Perhaps some foreigners will discover the work of the German reformers" (1954, p. 243). Dräger (1956, p. 5) wondered whether those of them who had emigrated (mentioning specifically G. Colm, A. Lowe, and H. Neisser) had not had an influence in shaping American policy under Roosevelt. The first significant

Proponents of an active countercyclical policy, in which public deficit spending would be an essential—but not necessarily the only—element could be found in various segments and strata of the Weimar Republic during the fateful years just prior to Hitler's seizure of power. Several academic economists—most in the junior ranks—civil servants, economic journalists, and others, made vigorous attempts in the early thirties to obtain public support for their views and to bring them to the attention of policymakers. Among them, at least two groups made an attempt to enlist the prestigious Keynes in support of their activities.[15]

These two groups centered on the Studiengesellschaft für Geldund Kreditwirtschaft (Society for the Study of Money and Credit), formed toward the end of 1931 by Wilhelm Grotkopp and Heinrich Dräger (the former, a business journalist, and the latter, a successful businessman, both holding Ph.D. degrees in economics) and W. S. Woytinsky, statistician for the German Trade Union Federation (ADGB), who had acquired a wide reputation as the author of the seven-volume *Die Welt in Zahlen* [The world in figures] (1925–28). The policies they advocated found practically no support among the established academicians, particularly among holders of chairs of political economy, with the notable exceptions

---

mention in English of the efforts of German activists, notably those by Woytinsky and Lautenbach, has recently become available in Kindleberger (1974), who makes numerous references to relevant memoir literature of those in power who opposed the activists' policy proposals and to other original sources. See pp. 173–77 for the first expansionary measures along the lines of Lautenbach's proposals following the fall of the Brüning government and prior to the advent of Hitler. Kindleberger errs, however, in saying that Woytinsky had presented a multiplier-mechanism analysis.

[14] While Keynes emphasized the interest rate, and his followers in the United States mostly stressed compensatory government spending, the German discussion was mainly in terms of the propriety and dangers, respectively, of "credit creation" to finance government investment expenditures. This is entirely understandable in view of the traumatic experience with runaway inflation still in everybody's mind and the widespread concern with protecting the integrity of the mark, which some of the reformers believed had assumed proportions of "reserve-cover mania."

[15] Later, Keynes had numerous contacts with refugee economists from Germany, of whom there were a number in England and some of whom, at least, were familiar to various degrees with the countercyclical policy debates. Thus, for instance, Jacob Marschak, who between 1933 and 1939 was reader in statistics and director of the Institute of Statistics at the University of Oxford, was the author of "Some Theses on Crisis Policies" (1931), which appeared in a well-known economic weekly published jointly by the University of Hamburg and the prestigious World Economic Institute of the University of Kiel. That article had been circulated earlier as a memorandum among some groups interested in providing an alternative to Brüning's deflationary policies. The article favored international action for public works to create jobs to be financed by credit expansion. The supporting argument is close to that of *General Theory*. Marschak, who had studied under Lederer and was a close friend of Woytinsky, was then a lecturer (*Dozent*) at Heidelberg University, where, incidentally, during summer 1931 he gave a seminar on Keynes.

of Adolph Lowe, Emil Lederer, and Werner Sombart. Some junior members of the academic community, however, such as F. Baade, G. Colm, L. Erhard (the post–World War II minister of the economy), J. Marschak, and H. Neisser favored them in one form or another. Even in the bureaucracy, there were supporters of credit expansion as a means of an activist countercyclical policy, such as Ernst Wagemann (who simultaneously headed the Reich's Statistical Office and the Institute of Business Cycle Research), who, early in 1932, published a plan for combating the crisis to which his name became attached,[16] and Wilhelm Lautenbach,[17] an official in the Ministry of Economics who had, perhaps, the clearest overall view of the problem. The reflationary proposals were opposed by the most influential German economic and business weekly, the *Deutsche Volkswirt*, published by Gustav Stolper,[18] and by all the political parties of the Weimer Republic.

The society for the Study of Money and Credit held during 1932 a number of meetings at which various proponents of an active countercyclical policy (Schumpeter canceled out because of his impending departure for the United States) gave papers. Copies of some of these papers, as well as the book on creating jobs by one of the society's main sponsors, Dräger, were sent to interested individuals, including Keynes. Keynes was asked early in 1932 to come to Berlin to give a talk to the society, but declined the invitation.[19]

Woytinsky launched his campaign from the solid base of the trade union movement, whose members had felt the full impact of the Depression. After completing his monumental statistical compilation. Woytinsky in 1929 became the head of the statistical office of the Trade Union Federation. Originally his efforts had been confined to producing better statistics on collective bargaining agreements as well as in other areas of labor statistics. But throughout his life Woytinsky had been an activist, and his creative mind could not be satisfied with improving statistics alone. He became intrigued by the causes of the worldwide decline in prices which was one of the justifications in Germany for deflationary policies. The result was *The International Raising of Prices as a Way Out of the Crisis* (1931*a*). It argued for price reflation—a policy goal which he soon dropped—and contained an articulated plan for public works on an international scale, to be financed by central bank credit.[20] In June 1931, Woytinsky, having become convinced of the futility of an international approach, published in the trade union monthly an article advocating credit-financed public works for Germany (1931*b*).

---

[16] The 66-page "memorandum" (Wagemann 1932) was the result of several months' discussions by a group of officials of the Statistical Office of the Reich and of the Institute for Business Cycle Research, as well as by experts connected with banking, industry, and academia. While in the text there is reference to the similarity between the proposals of the memorandum and those of the Report of the Macmillan Committee (which had been just published and of which Keynes was a leading member), there is no direct reference to Keynes.

[17] His pertinent publications are included in the collection of essays published after the war (Lautenbach 1952).

[18] Because (in the words of his widow and close collaborator) "with the continuing world crisis and without coordination with foreign countries, a one-sided German credit expansion would spend itself in additional capital flight and weakened currency and no positive result would remain" (Stolper 1960, p. 289). Another close collaborator of Stolper wrote many years later to Woytinsky's widow that in the "controversy I had with him on employment policy in the *Deutsche Volkswirt* in 1932, retrospectively I have come to the conclusion that he had the better case" (Carl Landauer to E. S. Woytinsky, July 12, 1961).

[19] D. E. Moggridge found in the Keynes correspondence files a copy of a letter declining the invitation (personal communication, December 22, 1973). In any case the meeting was canceled at the request of the German authorities, who feared that Keynes might repeat the recommendation to devalue the mark which he had made in an address delivered in Hamburg following the pound devaluation in September 1931 (see Dräger 1956, p. 98).

Fully aware of the limited public impact to be expected of any book published under the auspices of a research institute, Woytinsky started a publicity campaign for his plan.[21] For this purpose, he enlisted one of the most able members of the board of the Trade Union Federation, Fritz Tarnow, who several years earlier had published a book favoring expansionary policies, and Fritz Baade, then head of the government agency for agricultural marketing, who had academic connections and was a former member of the staff of the Research Institute (see below). To popularize his proposals, which became known as the "WTB Plan" (from the initials of its sponsors), Woytinsky published in 1931 and 1932 a number of articles in various media, ranging from trade union publications to newspapers and economic weeklies. To attract international attention, he published an article explaining the Plan in the quarterly of the International Labor Office and in two periodicals championing a European union. Woytinsky also attempted to enlist the support of Keynes, by addressing to him the following letter, which I have translated from the German.

---

[20] Woytinsky's statements in his memoirs that at that time he was not familiar with "the early works of Keynes" (1961, p. 464) must have been a lapse of memory, since he quoted Keynes several times.

[21] A major study of the campaign to use deficit-financed public works as a means of overcoming the Depression, including Woytinsky's role in it (M. Schneider, *Das Arbeitsbeschaftungsprogramm des Allgemeinen Deutschen Gewerkschaftsbundes*), is to be published under the sponsorship of the Friedrich Ebert Stiftung of Bonn. The essential facts can be found in Woytinsky's memoirs and in "Fighting Depression in Germany" by F. Baade, the only surviving sponsor of the Plan (in Woytinsky 1962). In a speech in the Reichstag on May 11, 1932, Theodor Heuss, the future first president of the Bundesrepublik, in referring to a speech by Gregor Strasser, the leader of the left wing of the National Socialists, referred mockingly to their "beautiful catchword"—"credit creation for productive purposes"—and drew attention to its similarity with the policies advocated by Woytinsky. Heuss claimed that anybody with a little grounding in economic theory would recognize them as a mixture of German romanticism and early utopian socialism. See Heuss (1963, pp. 388–89).

December 4, 1931

Esteemed Professor!

I have the honor of sending you, under separate cover, my book "The International Raising of Prices as a Way Out of the Crisis." Following a different path, I have reached essentially the same conclusions with regard to economic policy as those which you have been advocating in your country with such admirable energy, and I have had more than one opportunity to refer in my book to your investigations and views.

I have succeeded in winning for these views the support of the leading circles of the German trade unions, and in November the General Trade Union Congress had placed this issue before the Board of the International Trade Unions Federation under the heading of "International Job Creation on the Basis of Credit Creation." The plan sketched out in my book on pages 136/137 will be presently examined by an international Committee of Experts.

Mr. Citrine[22] told me that you might find it possible to participate in the Committee's work as an English expert. This would be a big success for the cause. The deflationary psychosis is still very strong in Germany, and sound judgment is emerging only slowly and painfully. Even in the ranks of the labor movement, which obviously cannot expect anything from a revaluation, one encounters at every step deflationary doctrines. I have tried in my book to free active economic policy from the mysticism of gold.

I would be very glad if you could find an opportunity to give attention to my study. But I would be even more pleased if I could meet you and make your personal acquaintance in the Committee on Experts of the International Trade Union Federation.

Very truly yours,
W. S. Woytinsky

It is not clear what authority Woytinsky had to extend an invitation to Keynes. The committee to which Woytinsky's letter alludes was never appointed. The Board of the International Labor Federation did not accede to the request of the German unions to place Woytinsky's proposals on the agenda of a forthcoming meeting. One can also wonder why Woytinsky expected to be able to interest Keynes in working on a trade union project, since Keynes's alignment with the Liberal party reflected not only espousal of Lloyd George's economic policies but also a rejection of the Labor party because it was "a class party, and the class is not my class."[23]

---

[22] The head of the British Trade Union Congress.

Keynes's answer was, as might be expected, negative:

> Bloomsbury
> 10th December 1931
>
> Dear Mr. Woytinsky,
>
> I am very much obliged to you for your letter of December 4, and for kindly sending me a copy of your new book, which I shall look through with great interest.
>
> I am, of course, in much sympathy with you in combatting the deflationary psychosis; it does seem to me in the degree to which it has now developed in Germany, to be a fearful thing. But I have much pressure of other work, and my knowledge of spoken German is not what it should be; so I am afraid that I cannot take part in the discussions to which you refer.
>
> I am sending you, under separate cover, a book of my own collected essays [*Essays in Persuasion*], which has recently appeared, which includes amongst other things some essays written for popular consumption, against the deflationist tendencies in this country.
>
> Yours very truly,
> J. M. Keynes

In referring to "spoken German" and to the deflationary psychosis in Germany, the reply disregarded the fact that a committee on an international plane was involved.

Subsequently, Woytinsky reformulated his proposals, adapting the Plan to conditions then prevailing in Germany and dropping its international aspects. The revised Plan drew fire from conservative economists of various political persuasions. The Social-Democrats, on whose support Woytinsky counted in the first place, refused to support it. Those in the Social Democratic party who opposed the WTB Plan were led by Rudolf Hilferding, its leading Marxist theoretician, the editor of its theoretical monthly *Die Gesellschaft,* and a former minister of finance (who later perished in the hands of the Gestapo). F. Naphtali, the head of the Research Institute (Forschungsstelle), jointly financed by the Social Democratic party, the trade unions, and consumer cooperatives, was another leading opponent in the socialist ranks. The modified Plan was, however, formally adopted by the Trade Union Federation in April 1932 and became its Plan to Create Jobs. The attempt of the German Trade union leaders, nearly all of whom were Social Democrats, to obtain for it the support of the Strasser wing of the Nazis is one of the most interesting episodes discussed and documented in Schneider's forthcoming book (see n. 21 above).

---

[23] For more on this, see "Political Liberalism, Class Loyalties and Anti-Marxism" (Lekachman 1966, p. 49).

More than 20 years after the conquest of power by Hitler, several political leaders and central bankers who were responsible for deflationary policies and who had survived, in Germany or in exile, claimed that they always had been reflationists at heart. Ex-Chancellor Brüning, whose deflationary policy had contributed significantly to the rise of Nazism, claimed that he (Brüning) had favored an expansionary credit policy and a public works program, but that provisions of the Young Reparations Plan (and the need to secure approval from the Bank for International Settlements, its agent) constituted insurmountable hurdles for implementing such a policy.[24] Brüning indicated that, while continuing to feel hamstrung by the negotiations on the cancellation of reparations that were in progress at the time, his government began to formulate as early as June 1930 a public works program, and that it also had even initiated secret negotiations with Gregor Strasser, the leader of the left wing of the Nazis. The then head of the Reichsbank, Hans Luther, asserted that the central bank could not endorse openly anything that would create in the population a feeling of uncertainty about the stability of the currency.[25] He furthermore claimed that, in fact, in fall 1932 the credit policy of the Reichsbank was expansionary.

---

[24] Brüning to Dräger, September 13, 1954 (see Brüning 1970, p. 137). In his memoirs Brüning does not discuss or justify his domestic economic policy. He mentions, however (p. 506), that he had "attempted, in a lengthy conversation [January 11, 1932] to convince Professor Keynes that in advocating inflationary techniques he was undermining any kind of reasonable financial policy for Germany. The audience which heard his Hamburg address had mistakenly concluded that the British Government shared his views. When I asked him how he thought [in applying his policy recommendation to Germany] of getting around the stipulations of the Young Plan with regard to the German currency, it turned out that he had not considered them at all." Hjalmar Schacht, who, until resigning in March 1930, was president of the Reichsbank, does not mention in his memoirs any of the attempts in favor of an active employment policy. He explained his willingness to resume the presidency under Hitler, on the explicit condition that the Reichsbank would provide credit to reduce unemployment (actually to finance rearmament), by his "resolve to risk the attempt, under a strong Government, to stem the tide of economic misery and provide wages and food once again to the six and a half million unemployed" (Schacht 1955, pp. 296, 302–3).

[25] See chapter 14 in Luther (1964). It contains, in addition to reasons for Luther's rejection of the recommendations of the monetary activists, the claim that, in fact, the policy of the Reichsbank had veered in the direction of expansion, and a quotation from a speech by Edgar Salin, the secretary of the Friedrich-List-Gesellschaft, who warned about the effect on the mark exchange rate of adopting policies which paralleled those of Keynes and were rejected in his own country. Luther reiterated this point in a letter dated April 2, 1954 (cited in Dräger 1956, p. 154). The same claim was also made by Vocke, at that time a high Reichsbank official, and in the early sixties president of the Bundesbank (see Vocke 1973). See also Gerschenkron (1970, pp. 16–17, n. 2).

By the end of May 1932, the Brüning government had finally formulated a program of job creation, but a few days after discussing it with representatives of the foreign press, Brüning was forced by President Hindenburg to resign (May 31, 1932). A quick succession of events made Hitler, half a year later, chancellor of the Reich and it was he who put into effect, by the means of a massive rearmament program, the basic ideas of those of his opponents who saw in an active job-creating countercyclical credit policy the only way for preserving German democracy.

No more than Roosevelt did Hitler have to await the publication of *General Theory* to embark on expansionary policies, even though the German edition followed the English original by only a few months. G. Strasser, G. Feder, and others in Hitler's party had already offered the prescription. Yet, in the foreword to the German edition of *General Theory* (dated September 7, 1936), after a few condescending remarks centering on the statement that "Germany . . . has been content for a whole century to do without any formal theory of economics which was predominant and generally accepted," Keynes offered his *opus* as manna: "How hungry and thirsty German economists must feel after having lived all these years without one!" Having pointed out that his analysis was developed mostly on the basis of conditions prevailing in the Anglo-Saxon countries, Keynes acknowledged that he was offering it to readers in Nazi Germany, in a curious statement for which there is no counterpart in his writings published in English: "Nevertheless, the theory of output as a whole, which is what the following book purports to provide, is much more easily adapted to the conditions of a totalitarian state, than is the theory of production and distribution of a given output produced under the conditions of free competition and a large measure of laissez-faire" (1973, 7:xxvi). In fact, as Joan Robinson pointed out nearly three decades later, "Hitler had already found how to cure unemployment before Keynes had finished explaining why it occurred" (1972, p. 8).[26] And, in addressing the 1972 Annual Meeting of the British Association for the Advancement of Science, Joan Robinson betrayed no more curiosity about the foreign forerunners of Keynes's policy recommendations than at the time she was one of Keynes's closest associates. She merely bemoaned the fact that "meanwhile the Nazis had been proving Lloyd George's point with a vengeance. It was a joke in Germany that Hitler was planning to give employment in straightening the Crooked Lake, painting the Black Forest white and putting down linoleum in the Polish Corridor" (1974).[27]

---

[26] A preceding sentence reads: "I do not regard the Keynesian revolution a great intellectual triumph. On the contrary, it was a tragedy because it came so late." Robinson's reassessment also came late. She first became widely known as the author of one of the earliest and most lucid expositions of the basic ideas of *General Theory* before it appeared.

[27] A much attenuated version appears in Robinson (1973).

Unfortunately, German jokes seem to have been better known in Cambridge than the theoretical and policy-oriented work of those German economists who had tried to develop policies that might have prevented the despair that ultimately paved the way for Nazism. In retrospect, the failure of the German reformers to persuade policymakers appears as one of the many tragic developments which contributed to the downfall of the Weimar Republic.

In the Economic Task Force's report to President-elect Kennedy, Samuelson (1962) wrote: "History reminds us that even in the worst days of the great depression there was never a shortage of experts to warn against all curative public actions, on the ground that they were likely to create a problem of inflation. Had this counsel prevailed here, as it did in pre-Hitler Germany, the very existence of our form of government could be at stake. No modern government will make that mistake again."

References

Ayres, C. E. "The Impact of the Great Depression on Economic Thinking." *A.E.R. Papers and Proc.*, vol. 36 (May 1946).

Brüning, Heinrich. *Memoiren, 1918–1934.* Stuttgart: Deutsche Verlagsanstalt, 1970.

Copeland, Morris A. "Economic Theory and Natural Science Theory." *A.E.R.*, vol. 21 (March 1931).

Davis, J. Ronnie. *The New Economics and Old Economists.* Ames: Iowa State Univ. Press, 1971.

Dillard, Dudley. "Gesell's Monetary Theory of Social Reform." *A.E.R.*, vol. 32 (June 1942).

Dräger, Heinrich. *Arbeitsbeschaftung durch produktive Kreditschöpfung.* Düsseldorf: Econ-Verlag, 1956 (first published in 1932).

Eccles, Marriner S. *Beckoning Frontiers.* New York: Knopf, 1951.

Gaitskell, H. T. N. "Four Monetary Heretics." In *What Everybody Wants to Know about Money*, edited by C. D. H. Cole. New York: Knopf, 1933.

Gerschenkron, Alexander. "Reflections on European Socialism." In *Essays in Socialism and Planning: In Honor of Carl Landauer*, edited by Gregory Grossman. New York: Prentice-Hall, 1970.

Grotkopp, Wilhelm. *Die grosse Krise.* Düsseldorf: Econ-Verlag, 1954.

Hansen, Alvin H. *Business Cycle Theory.* New York: Ginn, 1927.

———. *Business Cycles and National Income.* New York: Norton, 1951.

———. "The Influence of Keynesian Thinking in the United States." *Weltwirtschaftliches Archiv*, vol. 69, no. 1 (1952).

Harris, Seymour E., ed. *The New Economics: Keynes' Influence on Theory and Public Policy.* New York: Knopf, 1947.

———. *John Maynard Keynes.* New York: Scribner's, 1955.

Heuss, Theodor. *Erinnerungen.* Tübingen: Wunderlich, 1963.

Keynes, J. M. *Treatise on Money.* New York: Harcourt Brace, 1930.

———. *The Collected Writings.* Edited by Donald Moggridge. Vol. 7. *The General Theory of Employment, Interest, and Money.* Vols. 13, 14. *The General Theory and After*, pts. 1, 2. London: Macmillan, 1973.

Kindleberger, Charles. *The World in Depression, 1929–1939.* Berkeley: Univ. California Press, 1974.
Lautenbach, Wilhelm. *Zins, Kredit, und Produktion.* Berlin, 1952.
Lekachman, Robert. *The Age of Keynes.* New York: Random House, 1966.
Luther, Hans. *Vor dem Abgrund, 1930–1933.* Berlin: Propyläen, 1964.
Marschak, Jacob. "Einige Thesen zur Krisenpolitik." *Wirtschaftsdienst,* December 18, 1931.
Myrdal, Gunnar. *Against the Stream: Critical Essays on Economics.* New York: Pantheon, 1973.
Neisser, Hans. "Keynes as an Economist." *Social Research,* vol. 13 (June 1946).
Robinson, Joan. "The Second Crisis of Economic Theory." *A.E.R. Papers and Proc.,* vol. 62 (May 1972).
———, ed. *After Keynes.* Oxford: Blackwell, 1973.
———. "What Has Become of the Keynesian Revolution?" *Challenge,* vol. 16 (January-February 1974).
Samuelson, Paul A. "Prospects and Policies for the 1961 American Economy." *Taxation and Finance,* January 6, 1962.
Schacht, Hjalmar. *My First Seventy-six Years.* London: Wingate, 1955.
Schumpeter, J. A. Review of *The General Theory of Employment, Interest, and Money. J. American Statis. Assoc.,* vol. 31 (December 1936).
———. *History of Economic Analysis.* New York: Oxford Univ. Press, 1954.
Stein, Herbert. *The Fiscal Revolution in America.* Chicago: Univ. Chicago Press, 1969.
Stolper, Toni. *Ein Leben in Brennpunkten unserer Zeit.* Tübingen: Wunderlich, 1960.
Sweezy, Alan. "The Keynesians and Government Policy, 1933–1939." *A.E.R. Papers and Proc.,* vol. 62 (May 1972).
Trescott, Paul B. "Review of Davis' *The New Economics and Old Economists.*" *J. Money, Credit, and Banking,* vol. 5 (May 1973).
Tugwell, Rexford G. *The Democratic Roosevelt.* Baltimore: Pelican, 1969.
Vocke, Wilhelm. *Memoiren.* Stuttgart: Deutsche Verlagsanstalt, 1973.
Wagemann, Ernst. *Geld- und Kredit Reform.* Berlin, 1932.
Woytinsky, E. S., ed. *So Much Alive.* New York: Vanguard, 1962.
Woytinsky, W. S. *Die Welt in Zahlen.* 7 vols. Berlin: Mosse, 1925–28.
———. *Die internationale Hebung der Preise als ein Ausweg aus der Krise.* Leipzig: Buske, 1931. (*a*)
———. "Aktive Weltwirtschaftspolitik." *Arbeit,* vol. 6 (June 1931). (*b*)
———. *Stormy Passage.* New York: Vanguard, 1961.

More on Windfall Income and Consumption

Tong Hun Lee

University of Wisconsin—Milwaukee

I. Introduction

The permanent-income hypothesis (PIH) of consumption postulates that the marginal propensity to consume (MPC) of transitory income is zero. It maintains that, since windfall receipts are largely transitory income, they tend to increase consumption only insofar as they increase the permanent or long-run expected income. This feature of the PIH has important implications. According to Friedman's original exposition, the zero MPC of transitory income "introduces important substantive content into the hypothesis" (1957, p. 27). Also, the zero MPC of transitory income contradicts the absolute-income hypothesis, which postulates a strong effect on consumption of windfall income as well as regular income. Moreover, it raises an important policy question of whether a temporary income tax cut may have a stimulating effect on consumption.

Originally Bodkin (1959) tested this feature of the PIH by utilizing the 1950 BLS Urban Consumption Survey, in which information on the National Service Life Insurance (NSLI) dividend payment to veterans was collected. Defining the dividend as a windfall by following Friedman's original suggestion,[1] Bodkin found that the estimate of the MPC of windfall income was quite large and, indeed, as large as the estimate of the MPC of regular income. Thus, his finding challenged the validity of the PIH. Friedman (1960), however, replied that, since Bodkin estimated the effect of windfall income without holding permanent income constant, his estimate of the windfall effect might be biased upward, reflecting the effect of permanent income in disguise. In support of this view, Reid

---

I wish to thank Professors Ronald G. Bodkin and Milton Friedman for their constructive comments. I am also grateful to Professors Guy H. Orcutt and Arnold Zellner for their earlier guidance which enabled me to initiate this research. However, I am alone responsible for any shortcomings. This work was supported by the National Science Foundation under grant GS-40073 and also in part by funds provided by the College of Letters and Science and the Graduate School of the University of Wisconsin—Milwaukee. The data used in the paper were generously made available to me by Professor Bodkin.

[1] Friedman originally made this suggestion by stating that "this [NSLI] payment seems to have been largely unexpected and so to qualify as a 'windfall' orclearly transitory element" (1957, p. 215).

[*Journal of Political Economy,* 1975, vol. 83, no. 2]
© 1975 by The University of Chicago. All rights reserved.

(1962, 1963) reported that, when the geographical-place average of income was used as a measure of permanent income in the analysis, the results did not show any effect of windfall income on consumption. Unfortunately, Reid's results were not strictly comparable to Bodkin's because she employed a different definition of windfall income and used grouped data rather than individual observations.[2] To extend Reid's analysis further, Bird and Bodkin (1965) reestimated the windfall effect on consumption from individual observations by defining the NSLI dividend as the measure of windfall income and also by holding the effect of permanent income constant. Contrary to Bodkin's original estimate, however, Bird and Bodkin obtained different results of the windfall effect. With consumer outlays defined as "net of durables" (consumption ideally defined from the viewpoint of the PIH), the MPC of windfall income turned out to be not much different from that predicted by the PIH. Thus, Bird and Bodkin conceded that Friedman's theory of the consumption function had not been rejected.

In this paper I shall first show that the test conducted by Bird and Bodkin was blurred by the presence of heteroscedasticity in the residuals. Then I shall remove the heteroscedasticity by employing a new estimation procedure and, as a result, shall provide more powerful tests of the effect of windfall income on consumption. My results will show that the effect of windfall income on consumption is significantly greater than that predicted by the PIH, a finding which is, to some extent, at variance with Friedman's theory of the consumption function.

II. Approach

In keeping with Bird and Bodkin, I express the ith household consumption, C_i, as a function of windfall income, d_i, permanent income, P_i (prior to the receipt of the windfall), and other variables z_{ij}, with $j = 3, 4, \ldots, K$, as follows:

$$C_i = \alpha_0 + \alpha_1 P_i + \alpha_2 d_i + \sum_{j=3}^{K} \alpha_j z_{ij} \qquad (i = 1, 2, \ldots, n). \qquad (1)$$

The unobserved permanent-income variable, in turn, can be specified in the following form:

$$P_i = \beta_0 + \beta_1 y_i + \beta_2 d_i + \sum_{j=3}^{K} \beta_j z_{ij} \qquad (i = 1, 2, \ldots, n), \qquad (2)$$

---

[2] Reid (1962) initially did not include the NSLI dividend in her measure of windfall gains and, later (1963), when she did include it, she added other payments connected with military service. Also, Bird and Bodkin (1965) showed that the use of group averages caused an aggregation bias in the estimates.

where permanent income is a function of its proxy variables including regular income, y_i (net of d_i). By inserting equation (2) in equation (1), we obtain the equation in observed variables:

$$C_i = (\alpha_0 + \alpha_1\beta_0) + \alpha_1\beta_1 y_i + (\alpha_2 + \alpha_1\beta_2)d_i + \sum_{j=3}^{K} (\alpha_j + \alpha_1\beta_j)z_{ij} + u_i \qquad (i = 1, 2, \ldots, n), \quad (3)$$

where u_i is added as a random disturbance term.

The PIH states that the MPC of permanent income, α_1, is about 0.9. Since the receipt of the veteran's dividend as a windfall tends to raise the stream of permanent income, the MPC of windfall income, α_2, is positive but less than α_1. According to the PIH, α_2 is about 0.3.[3] Following Bird and Bodkin, I assume that α_2 is closely approximated by the coefficient of d_i, $(\alpha_2 + \alpha_1\beta_2)$, in the above equation. This assumption seems to be quite reasonable. Since P_i is permanent income *prior* to the receipt of d_i, d_i is likely to be a poor proxy for permanent income and, hence, β_2 is likely to be close to zero. Moreover, when better proxy variables such as age, education, housing expenditures, and geographical-place variables are introduced as z_{ij}'s, β_2 tends to become even closer to zero and may even become zero.[4] Thus, α_2, the MPC of windfall income, is closely approximated by $(\alpha_2 + \alpha_1\beta_2)$.

For estimating purposes, I specify the variance of the disturbance term of equation (3) as proportional to the pth power of the expected value of consumption:

$$\text{Var}\ (u_i) = k[E(C_i)]^{\rho}, \quad (4)$$

where the zero value of ρ indicates homoscedasticity and a nonzero value of ρ indicates heteroscedasticity. Intuitively, one would expect that disturbances in higher consumption brackets are associated mainly with larger variable items, such as housing and other luxury consumption items. Thus, ρ is likely to be positive.

---

[3] If the time horizon of the typical household for permanent income is 3 years, as Friedman (1957, pp. 150–52) estimated, the stream of permanent income for 1 year is approximately one-third of the windfall income. Friedman, therefore, maintains that α_2 is nonzero and is about 0.3, which is one-third of the MPC of permanent income.

[4] The general notation z_{ij} for $j = 3, 4, \ldots, k$ stands for the specific variables: H_i, representing home ownership; B_i, house purchase in the survey year; A_i, age of the head of household; E_i, education of the head; L_i, index of the level of housing expenditures; H_iL_i, an interaction term; and the remaining 60 dummy variables denoting geographical places. For detailed descriptions of the variables and for the justification of the use of such variables, see Bird and Bodkin (1965). Two important points, however, should be reiterated. First, the above variables, together with regular income, comprise an extensive list of variables normally considered as proxies for permanent income and, in particular, include those variables used by Reid (regular income, housing expenditures, and geographical places). Second, when consumption includes housing expenditures, the use of housing expenditures as a regressor in the household consumption function tends to yield a downward bias on the estimated coefficient of windfall income. In other words, even if β_2 is not exactly equal to zero, that is, even if $(\alpha_2 + \alpha_1\beta_2)$ is somewhat greater than α_2, the likely downward bias in estimating $(\alpha_2 + \alpha_1\beta_2)$ would tend to offset an upward bias in approximating $(\alpha_2 + \alpha_1\beta_2)$ as α_2.

When the heteroscedasticity is present, we can transform equation (3) into the following equation:

$$\frac{C_i}{[E(C_i)]^{\rho/2}} = (\alpha_0 + \alpha_1\beta_0)\frac{1}{[E(C_i)]^{\rho/2}} + \alpha_1\beta_1\frac{y_i}{[E(C_i)]^{\rho/2}} + (\alpha_2 + \alpha_1\beta_2)\frac{d_i}{[E(C_i)]^{\rho/2}} + \sum_{j=3}^{K}(\alpha_j + \alpha_1\beta_j)\frac{z_{ij}}{[E(C_i)]^{\rho/2}} + v_i \quad (i = 1, 2, \ldots, n), \tag{5}$$

with the homoscedastic variance of the new disturbance term as

$$\mathrm{Var}\,(v_i) = \mathrm{Var}\left\{\frac{u_i}{[E(C_i)]^{\rho/2}}\right\} = k. \tag{6}$$

Differing from Bird and Bodkin, I estimate equation (5) by employing the weighted least-squares method suggested by Theil (1971).[5] Since ρ is not known, I employ the weighted least-squares method iteratively for $\rho = 0, 1, 2$, etc., and, by using an F test, choose the value of ρ that would yield the homoscedastic variance of the estimating equation.[6]

---

[5] The weighted least-squares method involves a two-step procedure: first, estimate equation (3) by the least-squares method, ignoring the heteroscedasticity, and obtain the predicted value of consumption, $\hat{C}_i$; and second, substitute $\hat{C}_i$ for $E(C_i)$ in equation (5) and, *if* ρ *is known*, apply the least-squares method to this equation.

[6] Initially, I assumed that $\rho = 0$, thus maintaining the null hypothesis that disturbances in the estimating equation (5) are homoscedastic with $\rho = 0$. The null hypothesis can be tested by utilizing the method developed by Goldfeld and Quandt (1965) in the following way. Calculate $\hat{C}_i$ from equation (5) by the weighted least-squares method with $\rho = 0$ (or equivalently from eq. [3] by the least squares) and then divide the sample into two sets of observations: the first set with the lower values of $\hat{C}_i$ and the second set with the higher values of $\hat{C}_i$, both relative to the median of $\hat{C}_i$. Then compute the weighted least squares with $\rho = 0$ for both sets of observations separately and calculate the ratio of the residual sums of squares as

$$\frac{\sum_{i\in B}\hat{v}_i^2}{\sum_{i\in A}\hat{v}_i^2}, \tag{7}$$

where A and B refer to the first and second sets, respectively. Goldfeld and Quandt show that this ratio has an F-distribution with $(n/2 - k)$ degrees of freedom, both in the numerator and in the denominator. If the null hypothesis is accepted, the weighted least-squares estimates with $\rho = 0$ (the least-squares estimates) can be accepted, and the estimation process will end with the first round of computation. On the other hand, if the null hypothesis is rejected, we make a new assumption that $\rho = 1$, thereby deriving a new null hypothesis that disturbances in the estimating eq. (5) are homoscedastic with

This estimation procedure provides the consistent estimator of the coefficients of the household consumption function.[7] One may note that if $\rho = 0$, this estimation method is reduced to the ordinary least-squares method which Bird and Bodkin employed in their study.

III. Results

The weighted least-squares regressions of the household consumption function, obtained by using the data of Bird and Bodkin (1965), are reported in table 1 (with consumption to be construed as including the purchase of durables, c_t) and in table 2 (with consumption excluding durables, c_t'). Since c_t differs conceptually from C_t in the preceding section —C_t being defined by the PIH as the sum of the value of services derived from the stock of durable goods and the expenditure on nondurables— the c_t regressions may not provide a strict measure of the MPC of windfall income postulated by the PIH. For a rough test, however, I assume the value of 0.3 as the hypothetical value of the MPC of windfall income in the c_t regression.[8] On the other hand, although c_t' underestimates C_t, the c_t' regressions can be adjusted so as to provide the strict test needed. For this adjustment, as suggested by Bird and Bodkin, I use the value of

---

$\rho = 1$. The next step is to recalculate the ratio in (7) by inserting the weighted least-squares residuals from the estimating equation with $\rho = 1$. This ratio asymptotically approaches an F-distribution to be shown later. If the F-statistic is not significantly different from unity, we accept the underlying hypothesis and choose the weighted least-squares estimates with $\rho = 1$. This process completes the second round of iterative computation. If the test shows that $\rho \neq 1$, we can proceed to the third round of computations using the weighted least-squares method with $\rho = 2$ and so on. It should be noted that, if disturbances of the estimating equation increase with the higher values of consumption, the ratio in (7) tends to converge to unity as the sequence of positive integers for ρ approaches the population parameter. Therefore, ρ is uniquely determined and so is the particular round of iteration. When the ratio in (7) is calculated by using the residuals from the weighted least squares with $\rho \geq 1$, it has a limiting F-distribution. Assume that u_t is normally distributed with mean 0 and variance $k[E(C_t)]^{\rho}$. Then $u_t/(\hat{C}^{\rho/2})$ has a limiting normal distribution with asymptotic mean 0 and asymptotic variance k, and hence $\sum \hat{v}_t^2$ has a limiting χ^2 distribution. The ratio (7) consisting of the limiting χ^2-variates, both in the numerator and in the denominator, yields a limiting F-distribution. The resulting test on the heteroscedasticity from this ratio is an extension of the Goldfeld-Quandt test to the case of a large sample.

[7] As shown by Theil (1971), if the true parameter of ρ is used, the weighted least-squares method will yield the consistent estimator. In the present case, we choose ρ in such a way that the weighted least-squares estimator is consistent. In other words, given the value of ρ which closely approximates the true parameter, our estimation procedure yields the consistent estimator of the consumption function.

[8] Since Friedman provides the estimate of 0.9 for the MPC of permanent income when consumption is defined as c_t, we can assume the value of 0.3 as the hypothetical value of the coefficient of d_t in the c_t regressions. Friedman (1957, pp. 28–29), however, maintains that, when consumption is defined to include the purchase of durables, the correlation between transitory income and consumption may be greater than zero. Therefore, the value of 0.3, assumed here, may be an understatement of the MPC of windfall income under the PIH.

TABLE 1

ESTIMATED CONSUMPTION FUNCTIONS WITH c_t AS THE DEPENDENT VARIABLE (INCLUDES DURABLE-GOODS EXPENDITURES)

Equation	y_t (1)	d_t (2)	H_t (3)	B_t (4)	A_t (5)	E_t (6)	L_t (7)	H_tL_t (8)	F-Ratio (9)	t-Score (10)	t-Score (11)
(1)[a]	.6950**	.6461**	362.3*	473.3**	−0.146	58.35	245.8**	−178.5**	2.94**	2.10*	.29
	(.0193)	(.1648)	(143.5)	(98.6)	(5.090)	(38.24)	(36.9)	(42.8)	[$\rho = 0$]		
(2)[a]	.7135**	.7402**	110.8	401.1**	1.196	74.37*	197.5**	−102.1**	1.24*	2.96**	−.26
	(.0221)	(.1485)	(122.9)	(96.2)	(4.416)	(32.72)	(32.4)	(39.6)	[$\rho = 2$]		
(3)[a]	.7109**	.7854**	88.2	379.4**	2.070	78.85**	196.6**	−97.2**	0.94	3.34**	−.51
	(.0223)	(.1452)	(118.4)	(96.5)	(4.249)	(31.56)	(31.2)	(38.8)	[$\rho = 3$]		
(1′)[b]	.6939**	.6688**	358.6*	476.4**	−0.291	56.10	244.3**	−177.1**	2.88**	2.28*	.30
	(.0192)	(.1621)	(143.7)	(98.6)	(5.15)	(38.32)	(37.0)	(42.9)	[$\rho = 0$]		
(2′)[b]	.7134**	.7518**	113.2	405.2**	1.390	73.99*	197.7**	−103.0**	1.25*	3.09**	−.52
	(.0221)	(.1463)	(122.9)	(96.3)	(4.416)	(32.67)	(32.4)	(39.6)	[$\rho = 2$]		
(3′)[b]	.7111**	.7929**	91.1	383.8**	2.252	78.59**	196.8**	−98.3**	0.95	3.44**	−.56
	(.0223)	(.1431)	(118.4)	(96.5)	(4.250)	(31.51)	(31.2)	(38.8)	[$\rho = 3$]		

NOTE.—The figures in parentheses are the standard errors of the regression coefficients. The regressors listed above cols. 1–8 are defined in Section II of the text. Although the 60 geographical dummy variables are also included as regressors, their coefficients are not reported above because of space limitation.

[a] With the original y_t and d_t.

[b] With y_t and d_t rounded off to the accuracy of c_t.

* Significant at 5 percent level.

** Significant at 1 percent level.

TABLE 2

ESTIMATED CONSUMPTION FUNCTIONS WITH c_t' AS THE DEPENDENT VARIABLE (EXCLUDES DURABLE-GOODS EXPENDITURES)

Equation	y_t (1)	d_t (2)	H_t (3)	B_t (4)	A_t (5)	E_t (6)	L_t (7)	H_tL_t (8)	F-Ratio (9)	t-Score (10)	t-Score (11)
(1)[a]	.4978**	.3830**	336.4**	241.7**	13.93**	80.47**	251.7**	−173.3**	3.22**	1.20	.97
	(.0137)	(.1176)	(102.4)	(70.4)	(3.63)	(27.29)	(26.3)	(30.5)	[$\rho = 0$]		
(2)[a]	.5001**	.4217**	193.1*	199.7**	13.50**	74.34**	225.4**	−136.6**	1.31**	1.75	.75
	(.0155)	(.1029)	(86.0)	(65.0)	(3.16)	(23.05)	(23.2)	(28.0)	[$\rho = 2$]		
(3)[a]	.4936**	.4550**	176.6*	176.4**	13.56**	75.12**	224.3**	−135.1**	1.12	2.19*	.42
	(.0156)	(.0974)	(79.5)	(63.9)	(2.96)	(21.41)	(21.6)	(26.9)	[$\rho = 3$]		
(1′)[b]	.4969**	.4055**	334.3**	243.9**	14.04**	79.43**	251.0**	−172.7**	3.18*	1.42	.78
	(.0137)	(.1158)	(102.6)	(70.4)	(3.68)	(27.36)	(26.4)	(30.6)	[$\rho = 0$]		
(2′)[b]	.4994**	.4347**	194.8*	201.8**	13.65**	73.82**	225.5**	−137.1**	1.26*	1.90*	.63
	(.0155)	(.1012)	(86.0)	(65.0)	(3.16)	(23.00)	(23.1)	(28.03)	[$\rho = 2$]		
(3′)[b]	.4960**	.4648**	178.9*	178.5**	13.69**	74.64**	224.5**	−135.7**	1.06	2.38**	.32
	(.0152)	(.0936)	(77.6)	(62.4)	(2.89)	(20.84)	(21.1)	(26.3)	[$\rho = 3$]		

NOTE.—The figures in parentheses are the standard errors of the regression coefficients. The regressors listed above colums 1-8 are defined in section II of the text. Although the 60 geographical dummy variables are also included as regressors, their coefficients are not reported above because of space limitation.

[a] With the original y_t and d_t.

[b] With y_t and d_t rounded off to the accuracy of c_t'.

* Significant at 5 percent level.

** Significant at 1 percent level.

0.242 as the hypothetical value of the MPC of windfall income in the c_t' regression.[9] The resulting t-statistics of the differences between the coefficients of the d_t variable and their respective hypothetical values are given in column 10 of both tables.

To analyze our tests of the PIH, we should first evaluate the estimates of Bird and Bodkin which are presented in the regression equations with $\rho = 0$ in (1) of each table. While the t-score of the difference between the windfall effect and the hypothetical MPC of windfall income is significant in the c_t equation, it is not significant in the c_t' equation. Assuming that the c_t' equation was a more appropriate test of the PIH, Bird and Bodkin concluded that "these results appear reasonably consistent with the strict PIH" (1965, p. 508). Note, however, that the equations estimated by Bird and Bodkin have the heteroscedastic variances of residuals, as shown by the highly significant F-ratios in column 9 of tables 1 and 2. We can therefore proceed with the weighted regression analysis using the higher value of ρ.

When the weighted regressions are computed with successively higher values of ρ, equations with $\rho = 3$ for c_t and c_t' in the respective tables give the homoscedastic variances of residuals, as indicated by insignificant F-ratios. The regression coefficient of d_t in the c_t' equation now becomes significantly greater than has been assumed by the PIH. Also, the regression coefficient of d_t in the c_t equation is now larger than the corresponding coefficient previously obtained by Bird and Bodkin. Therefore, the conclusion of Bird and Bodkin should be modified.

We can further improve our results by making an appropriate adjustment to the errors in the variables originally coded by the BLS survey. In that survey, individual consumption expenditures were recorded in hundreds of dollars, whereas the NSLI dividends were given in tens of dollars. Thus, the tens of dollars in the dividend figures do not correspond to those of the consumption figures. Without such correspondence, they serve as errors in the variables and, hence, yield a downward bias in the estimate of the MPC of windfall income. A similar problem of errors in the variables arises when we use the regular income figures that were also coded in tens of dollars in the *BLS* survey.[10] For these reasons, I have recomputed the weighted regression equations after rounding off the dividend and regular income figures to hundreds of dollars. The results are presented in the lower half of each table. The estimates of MPC of windfall income become larger in each equation of the recomputed regressions in comparison with the corresponding equation that is not adjusted for errors in the variables. In particular, from the c'_t equation in (3′) we find that the estimate of 0.465 for the MPC of windfall income is greater than the hypothetical value of 0.242 at the .01 level of significance. This means that the correlation between transitory income and consumption is positive, and we must thereby reject the original postulate of the PIH.

---

[9] The value of 0.242 is derived by multiplying 0.3 by the ratio of the mean of c_t' to the mean of c_t. Since c_t' is defined in terms of flows as in the PIH, the c_t' regressions would provide a more stringent test of the PIH than would the c_t regressions.

[10] Since the ratio between the variances of regular income and windfall income is 76 to 1, the variance of the true components of regular income is likely to be far greater than the variance of the true components of windfall income. As a result, a downward bias in the estimated MPC of regular income would be substantially smaller than a downward bias in the estimated MPC of windfall income. Indeed, the regression results in the tables show this tendency.

Since the effect of windfall income is greater than that assumed by the PIH, one may ask whether the windfall effect is as large as the effect of regular income. To facilitate an answer to this question, I present the *t*-statistics of the difference between the coefficients of regular and windfall incomes in column 11 of the tables. Since these statistics are not significantly different from zero, one might be tempted to accept the validity of the absolute income hypothesis, which states that the effect of income is the same, regardless of the source of income. However, there are some indications in my results that would favor the PIH rather than the absolute-income hypothesis. Under the PIH, if a sufficient number of proxies for permanent income were included in any equation, then one would expect that the residual variation in regular income would essentially constitute transitory income; hence, the coefficient of regular income would tend to be small and even close to the coefficient of windfall income. In other words, the results of the insignificant *t*-statistics in column 11 do not necessarily support the absolute-income hypothesis by themselves, but they may also be consistent with the PIH.[11] Moreover, in table 2, which is the more appropriate table for the PIH, the coefficient of windfall income is somewhat smaller than that of regular income. Furthermore, in table 1, where consumption is defined to include durable-goods expenditures, the coefficient of windfall income becomes larger than that of regular income. Indeed, the PIH states that since windfalls tend to be invested, they go into durables to a greater extent than regular income. Thus, the PIH is supported by the fact that the MPC of windfall income is lower than that of regular income when the consumption concept excludes durable goods, while the reverse is true when it includes durables. Nonetheless, it is also true that the MPC of windfall income is greater than that postulated by the PIH, thereby implying that the MPC of transitory income is positive. Therefore, the results in this paper tend to support a weak PIH in which the effect of transitory income on consumption is positive, but less than that of permanent income.

---

[11] Although regular income partly reflects permanent income, the acceptance of permanent income as the more theoretically relevant variable can be taken to mean the rejection of the absolute-income hypothesis. In terms of symbols, it means that β_1 is less than unity. If β_1 is, say, 0.684, the coefficient of regular income, $\alpha_1\beta_1$, corresponding to consumption defined as c'_t, is 0.496, which we obtained in the c'_t regression in (3′).

In summary, my findings can be stated in terms of the structural parameters that are implied by the final estimates of equation (3′) of table 2. By following Friedman's procedure for omitting other variables from the consumption function, the structural equations in the preceding section may be derived as

$$C_i = .9\,P_i + .642\,d_i \tag{8}$$

and

$$P_i = .684\,y_i + 246\,H_i + 245\,B_i + 18.9\,A_i + 103\,E_i + 309\,L_i + 187\,H_iL_i + \ldots \text{etc.}, \tag{9}$$

where i ranges over all household observations. With an assumed value of 0.9 for the effect of permanent income on consumption, the coefficient of windfall income is smaller than that of permanent income, while it is greater than 0.3—the value postulated by the PIH. Also, the permanent-income variable is partly explained by regular income and other variables such as home ownership, H_i; house purchase, B_i; age, A_i; education, E_i; level of housing expenditures, L_i; and the interaction between home ownership and housing expenditure level, H_iL_i.

IV. Conclusions

Many economists, including Friedman, believed that, since the 1950 NSLI dividend payment to veterans was a genuine windfall income, these data would provide almost a controlled experiment for testing the validity of the permanent-income hypothesis. Despite repeated analysis, however, the previous studies have not provided conclusive results. In order to move toward a definitive conclusion, I have employed a more efficient method of estimation and a more powerful test of the permanent-income hypothesis. My results show that the marginal propensity to consume of transitory income is not zero, thereby suggesting that the original version of the permanent income hypothesis may not be entirely correct. However, the marginal propensity to consume of transitory income is not large enough to substantiate the absolute-income hypothesis. Indeed, the results tend to support a modest revision of the permanent-income hypothesis in which consumption is assumed to be affected positively by transitory income, but not as much as by permanent income. Also, unlike the implication of the permanent-income hypothesis, my findings suggest that a temporary reduction in individual income taxes would have some effect on consumption—consumption being defined as the sum of the expenditure on nondurables and the value of services consumed out of durable-goods stock. Moreover, when consumption is defined in terms of expenditures on durables and nondurables, the marginal propensity to spend transitory income becomes large. In other words, when consumption is defined in terms of a more policy-oriented variable, the windfall income created by a temporary tax cut may have considerable effect on consumer expenditures.

References

Bird, R. C., and Bodkin, R. G. "The National Service Life-Insurance Dividend of 1950 and Consumption: A Further Test of the 'Strict' Permanent Income Hypothesis." *J.P.E.* 73, no. 5 (October 1965): 499–515.

Bodkin, R. G. "Windfall Income and Consumption." *A.E.R.* 49 (September 1959): 602–14.

Friedman, M. *A Theory of the Consumption Function.* Princeton, N.J.: Princeton Univ. Press, 1957.

———. "Comments." In *Proceedings of the Conference on Consumption and Saving,* edited by I. Friend and R. Jones. Philadelphia: Wharton School Finance and Commerce, Univ. Pennsylvania, 1960.

Goldfeld, S. M., and Quandt, R. E. "Some Tests for Homoscedasticity." *J. American Statis. Assoc.* 60 (June 1965): 539–47.

Reid, M. G. "Consumption, Savings, and Windfall Gains." *A.E.R.* 52 (September 1962): 728–37.

———. "Consumption, Savings and Windfall Gains: Reply." *A.E.R.* 53 (June 1963): 444–4[illegible]

Theil, H. *Principles of Econometrics.* New York: Wiley, 1971.

Benefit-Cost Shadow Pricing in Open Economies: An Alternative Approach

Robin Boadway
Queen's University

I. Introduction

Little and Mirrlees (1968) have argued persuasively that, when assessing projects in open economies, the public sector should use world prices as its shadow prices for both traded and nontraded goods and factors.[1] For the latter, the world price means the value of foreign exchange used up by devoting resources to producing a unit of the nontraded commodity at the expense of a traded (numeraire) commodity. This is called the "foreign exchange equivalent" value of nontraded commodities. The theoretical underpinnings for such a view have recently been presented by Dasgupta and Stiglitz (1974). Using a general-equilibrium model of an open economy facing fixed world prices, they argue that shadow prices should be world prices for traded commodities and the foreign exchange equivalent for nontraded commodities unless there is a government budget constraint, a savings constraint, or a quota, or unless the level of net imports of a commodity influences nonoptimally some control of the government. (These are their rules 2 and 7, p. 30.) This would appear to suggest that the existence of taxes or tariffs which may or may not be variable by the agency evaluating the project has no overt effect on the shadow-pricing rules.

The analytics of Dasgupta and Stiglitz are complete only for traded commodities. Unfortunately, they do not explicitly consider shadow pricing of nontraded commodities except when all taxes and tariffs are set optimally. When restrictions exist on taxes and/or tariffs, as is very likely to be the case, we are simply told that "the shadow price of a nontraded good is its marginal foreign exchange cost" (1974, p. 23). There is no proof of this proposition, nor is there any indication of how to obtain the marginal foreign exchange cost from market data.

The purpose of this paper is to suggest an alternate way to derive

The author would like to thank J. E. Stiglitz for helpful comments on an earlier draft of this paper.

[1] The arguments may be found in Little and Mirrlees (1968, chaps. 8 [secs. 8.3, 8.4], 9).

[*Journal of Political Economy*, 1975, vol. 83, no. 2]
© 1975 by The University of Chicago. All rights reserved.

shadow-pricing rules for open economies. The method will encompass both traded and nontraded commodities under any sorts of restrictions on taxes or tariffs. The method, of course, gives the same shadow-pricing rules for traded commodities as Dasgupta and Stiglitz (1974) and Little and Mirrlees (1968) obtained (i.e., world prices). However, for nontraded commodities, we find that the shadow prices depend on taxes and tariffs in a fairly complicated way unless *all* taxes and tariffs can be set optimally. If they can, the Dasgupta-Stiglitz results apply.

Two other results of interest will be derived. First, Harberger's weighted-average shadow-pricing rule will be shown to be applicable to nontraded commodities when cross-price effects are neglected. Second, in undertaking project evaluation, the planner must also evaluate the *welfare loss of changing any tax or tariff rate that may be required to* finance it. This loss will also depend upon all taxes and tariffs.

In all of this, we use the same model as Dasgupta and Stiglitz, simplified for expositional purposes. The next section recaps the relevant aspects of that model and shows the difficulties involved in analyzing the nontraded-commodities case with restrictions on taxes and tariffs. Following that, our alternative approach will be presented.

II. The Dasgupta-Stiglitz Economy with Restrictions on Taxes and Tariffs

The economy is a small open economy with a single consumer (to avoid distributional problems). The consumer maximizes a utility function $U(c)$ subject to a budget constraint $qc = M$, where c is the vector of commodities consumed (including factors supplied with a negative sign), q is the vector of consumer prices, and M is lump-sum income (explained below). Utility maximization yields a set of demand functions, homogeneous of degree zero:

$$c_i = c_i(q, M) \qquad i = 1, \ldots, n. \tag{1}$$

Consumer goods are supplied from three sources: net imports z_i, public production x_i, and private production y_i. Market clearing requires that

$$c_i = y_i + x_i + z_i \qquad i = 1, \ldots, n. \tag{2}$$

Private sector production is assumed efficient and constrained by the transformation surface:[2]

$$f(y) = 0. \tag{3}$$

The private sector faces a vector of prices p and maximizes profits py.

---

[2] Dasgupta and Stiglitz (1974) have used several private sectors. Since we are not concerned with taxing different sectors at different rates, we can treat the private sector as one sector facing a common vector of producer prices and operating efficiently. Our analysis could easily be extended to allow for differential taxes by sector.

This yields a set of supply functions:[3]

$$y_i = y_i(p) \qquad i = 1, \ldots, n. \tag{4}$$

The supply of net imports (demand for net exports) is perfectly elastic at given world prices r.[4] The balance-of-payments constraint is

$$rz = 0. \tag{5}$$

The public sector selects its vector of outputs (inputs are negative outputs) from the transformation surface:

$$G(x) = 0. \tag{6}$$

We assume production efficiency within the public sector.[5]

Finally, the relations between the three sets of prices are

$$q_i = p_i + T_i \qquad \text{all } i \tag{7}$$

and

$$p_i = r_i + t_i \qquad \text{all traded } i, \tag{8}$$

where T_i is the (per unit) rate of consumption tax, and t_i is the (per unit) tariff rate.[6]

In general, some z_i are zero (nontraded goods), some c_i are zero (intermediate goods), and some y_i are zero (not domestically produced). To keep things simple, assume there are no intermediate goods, and all traded goods are also domestically produced. There are n commodities domestically produced and consumed. Of these n, h are also traded. The extensions to include intermediate goods and goods not produced at home are straightforward.

There are three budget constraints in this economy. In addition to the consumer's budget constraint, there are budgets for the government and the private sector. They are

$$Tc + tz + px + t_\pi \pi = L \tag{9}$$

and

$$py = \pi, \tag{10}$$

---

[3] These supply functions require some further interpretation. For a constant-returns industry faced with fixed prices, inputs and outputs are indeterminate. Therefore, equation set (4) cannot be interpreted as "*ceteris paribus*" industry supply functions. Rather they are "general-equilibrium" supply functions showing how supplies are related to producer prices along a transformation surface. They are derived by maximizing national income at factor cost subject to the transformation constraint. Only $n - 1$ of (4) are independent, the remaining one being determined by the constraint (3). To avoid indeterminacies in (4), we assume that the transformation surface is strictly convex so (4) is single-valued.

[4] Dasgupta and Stiglitz set all $r_i = 1$ for simplicity. That procedure is not followed here for pedagogical reasons.

[5] For the justification for assuming production efficiency within the public sector, see Stiglitz and Dasgupta (1971). Notice that it does not imply production efficiency over the entire economy.

[6] We do not explicitly consider production taxes, since they are equivalent to combined consumption taxes and tariffs.

where Tc is tax revenue, tz is tariff revenue, px is public-sector profits (or losses), t_π is the tax rate on profits, π is profits, and L is the lump-sum payments or receipts of the consumer. Note that the consumer's lump-sum income, M, equals $L + \pi(1 - t_\pi)$. From Walras's Law we may deduce satisfaction of any one of the budget constraints from the rest and equation (5).

The solutions to this general-equilibrium system may be found by solving equations (1), (2), (4), (7), and (8) for the variables q, p, c, x, y, z, T, t. These are $4n + h$ in $6n + 2h$ unknowns. By Walras's Law only $4n + h - 1$ of these equations are independent. At the same time, because of the homogeneity of degree zero of demand and supply, we can normalize two sets of prices.[7] We arbitrarily set $q_1 = p_1 = 1$ (and $T_1 = 0$) and assume commodity 1 is traded. Thus there are $6n + 2h - 2$ variables. This leaves $2n + h - 1$ degrees of freedom which the government uses up by their choice of x, t, and $(n - 1)$ of the T. Mathematically, the same allocation can be arrived at by treating any $2n + h - 1$ of the variables as exogenous. Thus, Dasgupta and Stiglitz (1974) use x, z, and $(n - 1)$ of the q as controls in their analysis.

In this paper, we are interested in the case in which the project evaluator is not free to vary all taxes and tariffs. Suppose m of the n taxes are predetermined, as are all of the tariffs (for example). This reduces the number of endogenous variables in the system by $m + h$, leaving the authorities $2n - m - 1$ degrees of freedom. In practice they choose x and $(n - m - 1)$ of the t, but any $2n - m - 1$ variables could be treated as exogenous to give the same solution. So, x and $(n - m - 1)$ of the q could be chosen. However, the remaining m consumer prices are endogenous and therefore are functions of the controls.

To see how the Dasgupta-Stiglitz methodology would operate under these restrictions, we maximize an indirect utility function $V(q)$ with respect to the appropriate controls. Note that we are assuming no lump-sum income to the consumer here, implying either 100 percent profits tax or constant returns to scale. To simplify the interpretation of the results, we treat private production y as exogenous (rather than functions of p) and add the production constraint $f(y) = 0$ to the problem. Mathematically, there are now $3n - m - 1$ degrees of freedom which will be used up by the following controls: $(n)x$, $(h)z$, $(n - h)$ untraded y, and $(n - m - 1)$ tax-variable q. We maximize $V(q)$ subject to the constraints (2), (3), (5), and (6).

$$\begin{aligned} L = V(q) - \lambda \sum r_i z_i + \mu G(x) + \psi f(y) \\ + \sum p_k(y_k + x_k + z_k - c_k). \end{aligned} \tag{11}$$

---

[7] See Diamond and Mirrlees (1971) for a discussion of the normalizations involved.

Before maximizing (11), notice that, with fixed r and t, $dq_i = dT_i$ for traded goods. Also, $dq_j/dx_i = 0, j = 1, \ldots, n$, for all traded x_i. That is, any increase in public-sector use of traded commodities will come entirely at the expense of net imports and will not influence any prices.[8] Bearing these in mind, the first-order conditions are

$$\begin{aligned}\frac{\partial L}{\partial q_i} &= \frac{\partial V}{\partial q_i} + \sum_j \frac{\partial V}{\partial q_j}\frac{dq_j}{dq_i} - \sum_k \rho_k\left(\frac{\partial c_k}{\partial q_i} + \sum_j \frac{\partial c_k}{\partial q_j}\frac{dq_j}{dq_i}\right)\\ &= 0 \qquad \begin{cases} i \in \text{tax-variable} \\ j \in \text{nontraded fixed tax;}\end{cases}\end{aligned} \tag{12}$$

$$\begin{aligned}\frac{\partial L}{\partial y_i} &= \sum_j \frac{\partial V}{\partial q_j}\frac{dq_j}{dy_i} + \psi\frac{\partial f}{\partial y_i} + \rho_i - \sum_k\sum_j \rho_k \frac{\partial c_k}{\partial q_j}\frac{dq_j}{dy_i}\\ &= 0 \qquad \begin{cases} i \in \text{nontraded} \\ j \in \text{nontraded, fixed tax;}\end{cases}\end{aligned} \tag{13}$$

$$\begin{aligned}\frac{\partial L}{\partial x_i} &= \sum_j \frac{\partial V}{\partial q_j}\frac{dq_j}{dx_i} + \mu\frac{\partial G}{\partial x_i} + \rho_i - \sum_k\sum_j \rho_k \frac{\partial c_k}{\partial q_j}\frac{dq_j}{dx_i}\\ &= 0 \qquad \begin{cases} \text{all } i \\ j \in \text{nontraded, fixed tax;}\end{cases}\end{aligned} \tag{14}$$

$$\frac{\partial L}{\partial z_i} = -\lambda r_i + \rho_i = 0 \qquad i \in \text{traded.} \tag{15}$$

Equations (12) form the basis for deriving optimal tax rules. While we are not concerned here with these rules, it may be noted that the "proportionate reduction" rule cannot be derived from (12). Dasgupta and Stiglitz (1974) derive this rule for economies in which either all commodities are tradable or all taxes and tariffs are variable. The optimal tax rules obtained from (12) contain the additional terms dq_j/dq_i, which must be explicitly worked out.

Equations (14) give shadow-pricing rules for the private sector. For all x_i that are traded, $dq_j/dx_i = 0$; so (14) reduces to

$$\mu\frac{\partial G}{\partial x_i} + \rho_i = 0 \qquad i \in \text{traded.} \tag{14a}$$

Combined with (15), this gives

$$\frac{\partial G}{\partial x_i}\bigg/\frac{\partial G}{\partial x_1} = \frac{r_i}{r_1}. \tag{16}$$

---

[8] This will become apparent from eq. (22) below.

That is, for traded commodities, public-sector shadow prices should be world prices regardless of tax or tariff restrictions or the presence of nontraded commodities. This is because any increase in public sector use of them comes entirely from net imports (or outputs go as net exports) without disturbing allocation elsewhere.[9]

For nontraded commodities, dq_j/dx_i will not in general be zero. An increase in the use of x_i will directly affect its own market price and indirectly all other nontraded x_i prices through induced demand and supply changes. From (14), the shadow price of nontraded x_i is

$$\frac{\partial G}{\partial x_i}\Big/\frac{\partial G}{\partial x_1} = \left[\left(-\sum_j \frac{\partial V}{\partial q_j}\frac{dq_j}{dq_i}\right)\Big/\rho_1\right] - \frac{\rho_i}{\rho_1} + \sum_k \sum_j \frac{\rho_k}{\rho_1}\frac{\partial c_k}{\partial q_j}\frac{dq_j}{dx_i} \quad (17)$$

$$i \in \text{nontraded}.$$

Equation (17) has no easy intuitive interpretation. It involves Lagrangian multipliers which have no direct market price counterpart. If, as Dasgupta and Stiglitz suggest (1974, p. 23), it is the "marginal foreign exchange cost," it is unclear what this means in the present context; nor do they prove this contention in their paper. In the next section, we derive explicit shadow-pricing rules for the public sector which do not include Lagrangian multipliers. The method is sufficiently general to include any or no restrictions on taxes and tariffs. For example, the government need not be "optimizing" with respect to any taxes or tariffs but may simply use them to raise revenue in any arbitrary way.

III. Generalized Benefit-Cost Rules for Open Economies

To develop a general decision rule for the public sector, we take the general-equilibrium system formed by equations (1), (2), (4), (7), and (8); disturb the system by changes in x, T, and t; and see what happens to utility. To simplify matters, we assume no lump-sum incomes ($M = 0$) for consumers and constant returns ($py = 0$). To derive a measure for the change in utility, differentiate the utility function:

$$dU = \sum_i U_i dc_i.$$

From consumer behaviour we know $U_i/U_1 = q_i/q_1 = q_i$. Therefore

$$\begin{aligned} dU/U_1 &= dW = qdc \\ &= (p + T)(dy + dx + dz) \qquad (18) \\ &= Tdc + pdx + tdz, \end{aligned}$$

---

[9] Note also that this entails aggregate production inefficiency, since domestic producer prices do not equal world prices because of the tariff.

by using the fact that $pdy = 0$ from profit maximization,[10] and $rdz = 0$ from (5). This gives the welfare change in terms of the numeraire as a function of changes in c, x, and z. We would like to convert this to dW as a function of changes in those variables the government actually controls—T, t, and x. This is accomplished by using the reduced form of the general equilibrium system to solve for dc and dz in terms of dx, dT, and dt. Equation (18) may be rewritten (using [1] and $M = 0$) as

$$dW = T\frac{\partial c}{\partial q}dq + pdx + tdz. \tag{19}$$

To obtain expressions for the total effects of T, t, and x on q and z, we substitute (1), (4), (7), and (8) into (2). Distinguishing between traded and nontraded commodity vectors with the subscripts t and n, this yields the following n equations:

$$c_n(q_n, r + T_t + t_t) = y_n(q_n - T_n, r + t_t) + x_n;$$

$$c_t(q_n, r + T_t + t_t) = y_t(q_n - T_n, r + t_t) + x_t + z.$$

Total differentiation of these equations yields

$$c_{nn}dq_n + c_{nt}(dT_t + dt_t) = y_{nn}(dq_n - dT_n) + y_{nt}dt_t + dx_n; \tag{20}$$

$$c_{tn}dq_n + c_{tt}(dT_t + dt_t) = y_{tn}(dq_n - dT_n) + y_{tt}dt_t + dx_t + dz. \tag{21}$$

Of these n equations, only $n - 1$ are independent by Walras's Law. The equation for the numeraire good will be dropped. Using an asterisk to indicate that good 1 has been deleted from the vector or matrix in question, we can solve the remaining $n - 1$ equations for dq_n and dz^* as a function of dT^*, dt, and dx. From (20), the solution for dq_n is

$$\begin{aligned} dq_n &= (c_{nn} - y_{nn})^{-1}[dx_n - (c_{nt} - y_{nt})dt_t - c_{nt}dT_t - y_{nn}dT_n] \\ &= A_{nn}^{-1}(dx_n - A_{nt}dt_t - c_{nt}dT_t - y_{nn}dT_n), \end{aligned} \tag{22}$$

where $A_{nt} = (c_{nt} - y_{nt})$, the matrix of partial derivatives of domestic excess demands of nontraded goods with respect to traded goods' prices; A_{nn}, A_{tt}, and A_{tn} are similarly defined.

The solution for dz^* is obtained from (21) using (22):

$$\begin{aligned} dz^* = {} & -dx_t^* + A_{tn}^*A_{nn}^{-1}dx_n + (A_{tt}^* - A_{tn}^*A_{nn}^{-1}A_{nt})dt_t \\ & + (c_{tt}^* - A_{tn}^*A_{nn}^{-1}c_{nt})dT_t + (y_{tn}^* - A_{tn}^*A_{nn}^{-1}y_{nn})dT_n. \end{aligned} \tag{23}$$

---

[10] Proof: From producer efficiency in the private sector, $f(y) = 0$. Totally differentiating: $f_i dy_i = 0$. After normalizing, $f_i/f_1 dy_i = pdy = 0$.

Using these massive expressions for dz^* and dq_n in terms of government controls, and recalling that $dq_t = dT_t + dt_t$, (19) expands to

$$
\begin{aligned}
dW = {} & \left[T\frac{\partial c}{\partial q_t} - T\frac{\partial c}{\partial q_n} A_{nn}^{-1} c_{nt} + t(c_{tt}^* - A_{tn}^* A_{nn}^{-1} c_{nt}) \right] dT_t \\
& + \left[-T\frac{\partial c}{\partial q_n} A_{nn}^{-1} y_{nn} + t(y_{tn}^* - A_{tn}^* A_{nn}^{-1} y_{nn}) \right] dT_n \\
& + \left[T\frac{\partial c}{\partial q_t} - T\frac{\partial c}{\partial q_n} A_{nn}^{-1} A_{nt} - t(A_{tt}^* - A_{tn}^* A_{nn}^{-1} A_{nt}) \right] dt_t \\
& + r dx_t + \left[p + T\frac{\partial c}{\partial q_n} A_{nn}^{-1} + t A_{tn}^* A_{nn}^{-1} \right] dx_n .
\end{aligned}
\tag{24}
$$

This complicated expression is our generalized benefit-cost rule. It shows the change in welfare as a function of changes in taxes, tariffs, and public-sector outputs and inputs. Despite the complexity of (24), we can deduce the following results.

1. The public sector should use world prices r as shadow prices for its traded goods and factors. This follows directly from observing the weights on dx_t in the welfare measure (24).

2. Public-sector shadow prices for nontraded commodities are not interpreted as foreign exchange equivalents. Rather, they are the sum of producers' price and terms involving all T and t. To get a feel for the meaning of this, if we ignore all cross-price effects on both supply and demand, the shadow price of x_i reduces to

$$
p_i + \frac{T_i \partial c_i / \partial q_i}{\partial c_i / \partial q_i - \partial y_i / \partial p_i} = \frac{q_i \partial c_i / \partial q_i - p_i \partial y_i / \partial p_i}{\partial c_i / \partial q_i - \partial y_i / \partial p_i} . \tag{25}
$$

This is Harberger's weighted-average formula (see, e.g., Harberger 1969). It says public-sector shadow prices should be a weighted average of demand and supply prices, the weights being the proportions in which an increase in public sector-use of x_i comes at the expense of reduced private-sector demand and increased private-sector supply.

3. If the project evaluator is unable to vary any taxes or tariffs so T and t are fixed, and if we notice that $A_{nn}^{-1} = dq_n / dx_n$ (from [22]) and $A_{tn}^* A_{nn}^{-1} = dz^* / dx_n$ (from [23]), equation (24) becomes simply

$$
dW = r^* dx_t^* + \left[p + \frac{\partial (Tc)}{\partial q_n} \frac{dq_n}{dx_n} + \frac{d(t^* z^*)}{dx_n} \right] dx_n . \tag{26}
$$

The shadow price of a nontraded commodity is the private-sector producer price plus the change in total tax and tariff revenue resulting from a change in the output (or input) of the commodity in question by the public sector.

4. As a special case, suppose the government can choose all T and t. If it does so optimally, it makes $dW = 0$. It is more convenient to use the expression (19) for dW here. Analytically, the government chooses all x, T, and t to make $dW = 0$ subject to satisfying the budget constraint $Tc + px + tz = 0$. Totally differentiating the latter and rearranging yields[11]

$$\left(T^* \frac{\partial c^*}{\partial q^*} + c^*\right) dq^* + pdx + tdz = 0. \tag{27}$$

To make $dW = 0$ in (19) and at the same time satisfy (27), we set all tariffs t to zero, use p as shadow prices ($= r$ for traded goods) so that $pdx = 0$, and select T so as to satisfy the following equations:

$$T^* \frac{\partial c^*}{\partial q^*} dq^* = 0$$

and

$$\left(T^* \frac{\partial c^*}{\partial q^*} + c^*\right) dq^* = 0.$$

This can be solved for the optimal tax rules for the economy. These results on shadow pricing and optimal taxation correspond to the Dasgupta and Stiglitz "fully controlled" economy (1974, sec. 2) and to the Diamond and Mirrlees (1971) closed-economy results on production efficiency.

If, however, some T or t are predetermined, the shadow-price rules for x_n are not so straightforward. Suppose some of the T are predetermined. The government loses as many degrees of freedom as there are predetermined taxes. In (19), we could treat the dq_i of the fixed tax goods as endogenous and solve for them in terms of the rest of the dq and all dx_n (from the general-equilibrium system using a method similar to [22] and [23] above). The result would be an expression for dW in terms of the tax-variable dq, the dz, and the dx. The shadow prices of the dx,

---

[11] Totally differentiating the budget constraint gives

$$\begin{aligned} Tdc + cdT + pdx + xdp + tdz + zdt &= 0 \\ &= Tdc + c(dq - dp) + pdx + xdp + tdz + zdp \\ &= Tdc + cdq + pdx + tdz \\ &\qquad (\text{using } pdy = -ydp = 0) \\ &= (T^* \partial c^* / \partial q^* + c^*) dq^* + pdx + tdz \\ &\qquad (\text{using } T_1 = 0,\ dq_1 = 0). \end{aligned}$$

which are the coefficients in front of dx, would now include a term in all taxes T added to the producers' price.[12]

5. When taxes cannot be set optimally, one must take into consideration the welfare loss of any revenue requirements for a public project in addition to the shadow values of the commodities involved. These welfare costs are given by the terms in the square brackets in (24) in front of the tax and tariff changes. Once again, if we are prepared to ignore cross-price effects, this can be simplified considerably. For example, a tax increase on a traded good would have the following welfare cost:

$$dW = (T_i + t_i)\partial c_i/\partial q_i dT_i. \tag{28}$$

This is just the partial-equilibrium deadweight loss obtained from simple demand and supply curve analysis.

6. If producer prices for nontraded goods were fixed, then $dq_n = dT_n$ and $dq_t = dT_t + dt_t$. Using these in (19) and deriving an expression for dz as before, (24) becomes instead

$$dW = (T + t)\frac{\partial c}{\partial q}dT + (T + t)\frac{\partial c}{\partial q}dt - ty_{tt}dt_t + (p - t)dx. \tag{29}$$

The public-sector shadow prices are $p_i - t_i$, which, for traded goods, reduces to r_i. For nontraded goods, $t_i = 0$, so shadow prices become domestic producer prices. This reflects the fact that, with fixed supply prices, any increase in public-sector use comes entirely at the expense of private-sector supply, so the opportunity cost of the latter is applicable. Note also that as long as $t_i \neq 0$ for some i, aggregate production efficiency is not desirable.

The tax terms in (29) measure the deadweight loss from changing taxes and/or tariffs. For example, suppose we consider the effect of changing tariff t_i on welfare and ignore all cross-price effects. The deadweight loss may be written

$$(T_i + t_i)\frac{\partial c_i}{\partial q_i}dt_i - t_i\frac{\partial y_i}{\partial p_i}dt_i.$$

---

[12] So, for example, eq. (19) would read as follows (where f and g refer to commodities with fixed and variable taxes, respectively):

$$\begin{aligned} dW &= T\frac{\partial c}{\partial q_f}dq_f + T\frac{\partial c}{\partial q_g}dq_g + pdx + tdz \\ &= \left(T\frac{\partial c}{\partial q_f}\frac{dq_f}{dq_g} + T\frac{\partial c}{\partial q_g}\right)dq_g + \left(p_n + T\frac{\partial c}{\partial q_f}\frac{dq_f}{dx_n}\right)dx_n + pdx_t + tdz. \end{aligned}$$

The optimizing government sets $t = 0$, so $pdx_t = rdx_t$. The shadow prices of nontraded goods are the coefficients of dx_n. The optimal taxes are found by setting the variable T in such a way as to make the first term zero and at the same time satisfy the budget constraint of the government.

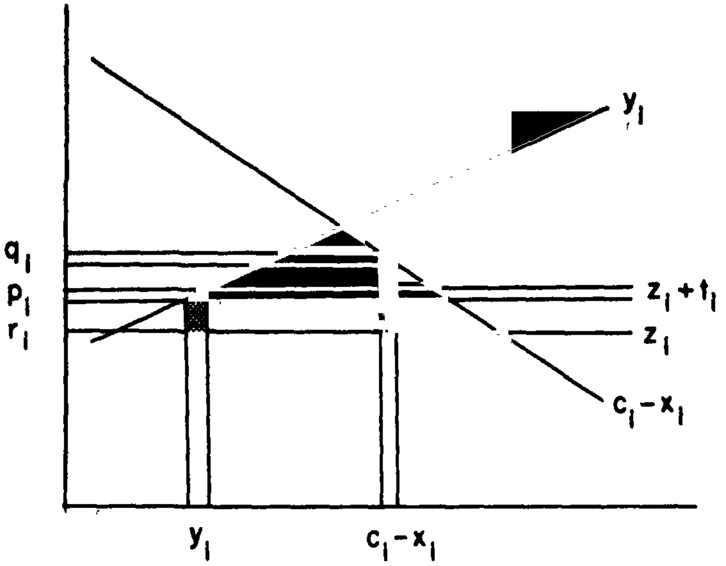

Fig. 1

This loss may be illustrated using supply and demand curves. Figure 1 shows the relevant supply curve, y_i; the demand curve net of government supplies (which are held constant), $c_i - x_i$; and the supply curve of net imports, z_i. The welfare loss from an incremental increase in the tariff is shown by the two shaded areas.

IV. Concluding Remarks

The use of world prices as shadow prices in the public sector is generally only valid for traded commodities in a tax- and tariff-distorted world. For nontraded goods, unless the government always sets taxes optimally and tariffs at zero, the shadow price depends on the existing set of taxes and tariffs. This may be of some importance for less developed countries where the level of tariff distortions is often quite high. A Harberger-type weighted-average formula may be the most appropriate shadow-pricing rule rather than the Little-Mirrlees "foreign exchange equivalent" for nontraded commodities.

Our model was developed using a number of simplifications, none of which is crucial. The distortions need not be due to taxes or tariffs. They could be due to monopoly, externalities, risk, etc. Public goods could be added with no loss of generality. The Dasgupta-Stiglitz constraints on savings, government, etc., could be added, as could intermediate goods and goods not produced domestically. Private-sector technology could be nonconstant returns. For example, with decreasing returns, pure profits are generated. If taxed away by the government, they would not affect equation (19) but would show up in the government budget constraint. If given to the consumer, they show up as a lump-sum income M and as an argument in the demand functions, thus affecting (19). These cases may be easily worked out by the reader.

It is also interesting to extend the model to a large open economy. That is, some world prices become endogenous to the system, implying that the country has some monopsony-monopoly power in international markets. To accommodate this, we simply need add an equation of the following sort to the system for each such good (or factor):

$$z_i = z_i(r). \tag{30}$$

In this case, the completely optimizing government sets the well-known optimal tariffs on those traded goods with variable world prices;[13] the shadow price of those goods then becomes the marginal foreign exchange cost (as Little and Mirrlees argued). However, as soon as we move away from the ideal planning case, shadow prices for those traded goods cease to be the marginal foreign exchange costs but depend on T and t in a fairly complicated way.

References

Boadway, R.; Maital, S.; and Prachowny, M. "Optimal Tariffs, Optimal Taxes, and Public Goods." *J. Public Econ.* 2 (November 1973): 391–403.

Dasgupta, P., and Stiglitz, J. E. "Benefit Cost Analysis and Trade Policies." *J.P.E.* 82, no. 1 (January/February 1974): 1–33.

Diamond, P. A., and Mirrlees, J. A. "Optimal Taxation and Public Production." *A.E.R.* 61 (March and June 1971): 8–27; 261–78.

Harberger, A. C. "Professor Arrow on the Social Discount Rate." In *Cost-Benefit Analysis of Manpower Policies*, edited by G. G. Somers and W. D. Wood. Kingston, Ont.: Industrial Relations Centre, Queen's University, 1969.

Little, I. M. D., and Mirrlees, J. A. *Manual of Industrial Project Analysis in Developing Countries.* Paris: Org. Econ. Cooperation and Development, 1968.

Stiglitz, J. E., and Dasgupta, P. "Differential Taxation, Public Goods and Economic Efficiency." *Rev. Econ. Studies* 38 (April 1971): 154–78.

---

[13] For a proof of this in an optimizing model, see Boadway, Maital, and Prachowny (1973).

On the Indirect Productivity of Money

Emil-Maria Claassen

University of Paris-Dauphine

The recent approach to the productivity of money in its medium-of-exchange function as it has been developed by authors like Brunner and Meltzer (1971) focuses on the amount of real resources allocated in the *exchange activity* of an economy; that is, in the purchase and sale of goods (commodities, assets, factors of production). The emergence of a money economy and a higher real quantity of money, respectively, increase the available amount of the "original" factors of production (capital, labor) for the proper *production activity*. Consequently, real money balances should be taken into account by the traditional production function. Either one treats them as a "third" factor of production as it is proposed by Mundell (1971, chap. 5) and Bailey (1971, pp. 54–56, 68–70), ascribing to money a "direct" productivity, or, as the present note suggests, they have only a "derived," "indirect" productivity, emphasizing the fact that money switches real resources from the exchange activity to the production activity. Even if both approaches may be considered as equivalent with respect to the result that money is productive, the latter constitutes a better analytical device because it reveals some new information on the individual's choice margin to hold money and on the specific shape of the demand curve for real money balances. Referring to a recent empirical study by Sinai and Stokes (1972) who show that, in a Cobb-Douglas production function, the presence of money is highly labor saving (the coefficient of capital services is hardly affected whereas the coefficient of labor services falls at about 20 percent, so the sum of the productive contributions of real balances plus labor is approximately equal to the coefficient of labor services in the production function without money), in what follows we shall assume that the amount of capital employed in the exchange activity is of a rather low magnitude, so we can neglect it.

The production function of the economy can be written as

$$y = y(\bar{K}, N^p), \tag{1}$$

[*Journal of Political Economy*, 1975, vol. 83, no. 2]
© 1975 by The University of Chicago. All rights reserved.

where y stands for real income or output (of the production sector), $\bar{K}$ stands for the existing amount of capital, and N^P is derived from the identity

$$\bar{N} = N^P + N^t, \tag{2}$$

indicating that the total amount of labor services ($\bar{N}$) available in the economy is used in two kinds of activities: the proper production activity for which parts of $\bar{N}$, that is N^P, are allocated and the trading activity requiring the input N^t. The output effect of money can be quantified as follows. (I) The transition of a barter economy to a money economy leads to a "quantum jump" in y because the trading activity in a money economy requires considerably less N^t and, hence, according to identity (2), makes available these resources for production. (II) Given the money economy, an increase in real balances (m) releases further amounts of labor,

$$N^t = N^t(m) \qquad N^t_m \leq 0, \tag{3}$$

for the use in the production process, so that according to (2)

$$N = N^P(m) \qquad N^P_m \geq 0. \tag{4}$$

By forming the derivative of y (eq. [1]) with respect to m in taking into account function (4), we can formulate clearly what can be understood by the "marginal productivity of money":

$$\frac{dy}{dm} = \frac{\partial y}{\partial N^P}\frac{\partial N^P}{\partial m}. \tag{5}$$

It is a derived, an "indirect" productivity: a change in the real quantity of money induces a change in real resources (labor) for production purposes, and only this latter leads to a change in output.

The upper right-hand panel of figure 1 traces the output effect of money, which is the outcome of the production function (1), illustrated in the upper left-hand panel as the traditional relationship with relatively decreasing returns, and of the "transformation function," (4), shown in the lower left-hand panel. (I) In the hypothetical case where $m = 0$, the amount of labor for production is supposed to be $\hat{N}^P$ producing the output $\hat{y}$. This value includes already the "quantum jump" in the quantity of labor available for production when the economy has moved from a barter economy to a monetary economy. (II) An increase in the real quantity of money from $m = 0$ to m_1 raises labor input in the production activity from $\hat{N}^P$ to N^P_1, inducing a gain in real output from $\hat{y}$ to y_1.

The fundamental problem with respect to the productivity of money and to its marginal value (5) is posed by the knowledge of the exact nature of the substitution function (3), giving rise to the transformation

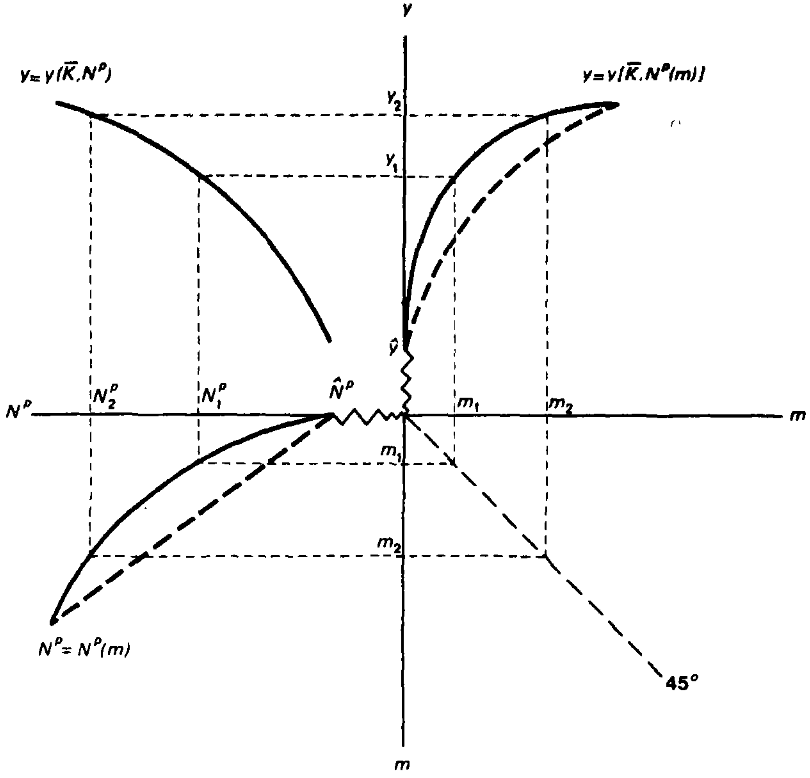

FIG. 1

function (4). In the lower left-hand panel of figure 1, two alternative relations have been traced. The most simple one is the linear one represented by the long-dashed straight line, which corresponds to the long-dashed productivity-of-money curve of the upper right-hand panel. The more sophisticated technological relationship is the nonlinear one which indicates decreasing savings of labor when real cash balances increase, so the productivity gain of a higher real quantity of money is especially significant if the increase takes place at a relatively low level of real cash balances. These two different interpretations of the "trading technology" result in the two alternative schedules for the marginal productivity of money in figure 2, where the solid (dashed) curve corresponds to the shape of the solid (dashed) curve of the total productivity of money in figure 1.

There is a great likelihood that the nonlinear case is the relevant one. Suppose that the transactions volume be T for the whole economy during a year. All transactions should employ money so that T represents the cumulative sum of money payments. Any payment can be effected in

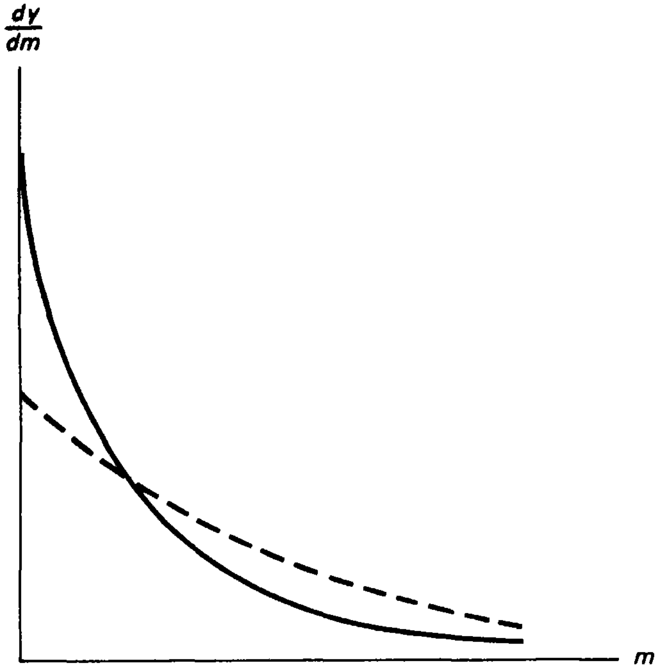

FIG. 2

small lots (S) which correspond to a higher number of transactions (n) and vice versa, because $T = nS$. Putting this expression on a strict macroeconomic level, it has to be identical with the equation of exchange where n is the velocity of money and S the real quantity of money such that $T = nm$. Suppose that the real cost of making a payment is solely a lump-sum amount, a, in terms of a constant amount of labor so that $N^t = an$. Because $n = T/m$ the substitution relation (3) can be reformulated more concretely as

$$N^t = a\frac{T}{m}, \tag{3a}$$

so that

$$N^p = \bar{N} - a\frac{T}{m}, \tag{4a}$$

which will have the shape shown by the solid curve in the lower left-hand panel of figure 1. Consequently, the marginal productivity of money can be written as

$$\frac{dy}{dm} = \frac{\partial y}{\partial N^p}\, a\frac{T}{m^2}. \tag{5a}$$

Because we treat money only in its function as a medium of exchange, the productivity of money is only the productivity of "transactions" and "precautionary" balances. The productivity of transactions balances arises from lower efforts to synchronize current receipts and expenditures, and the productivity of precautionary balances stems from the fact that otherwise there would be needed some amount of "idle labor" in order to be insured against the unforeseen variability in future transactions.

On which margin do individuals choose to hold money? Following the usual procedure of the return and cost of money holdings, let us turn first to the return component. An individual who decides to increase his money holdings by Δm will have a gain in terms of the less effort he has to spend in the activity of buying and selling goods. Thus, he has more time available either for his own leisure or for his proper professional activity. If he chooses to allocate his "free time" into his proper work, he will earn on his ΔN^P an additional income of $w \cdot \Delta N^P$, where w is the real wage rate. As is well known, he will undertake this decision on the leisure-work margin according to which he is only disposed to work if the utility of the additional income is higher than the satisfaction he would get from additional leisure. In the case that he chooses higher leisure instead of higher income, the return of money will have to be expressed in utility terms.

Suppose that the individual prefers to gain the additional income $w \cdot \Delta N^P$ (as we have assumed since the beginning of our study), which he would earn if he increases his cash balances by Δm. However, his definite decision will depend on the cost of Δm as well. These costs are opportunity costs equal to the return he can earn on alternative assets in comparison with non-interest-bearing cash holdings. If the rate of return on these alternative assets is r, the cost of Δm will be $r \cdot \Delta m$. Consequently, our individual will undertake an increase in cash balances so long as $w \cdot \Delta N^P > r \cdot \Delta m$, and his desired level of cash balances will be attained when the marginal return is equal to the marginal cost of holding cash, that is, when

$$w \frac{\Delta N^P}{\Delta m} = r. \qquad (6)$$

The left-hand expression represents another formulation of equations (5) and (5a) to the extent that w equals $\partial y/\partial N^P$.

Two results with respect to the demand-for-money function can be derived from the above analysis. The first refers to the real wage rate, which is one important variable affecting the demand for money. A higher w raises the return component of money holdings; raises the marginal productivity of money so that the demand for real balances will be an increasing function of w. Thus, in economies where the real wage rate is very high, the relative size of existing real cash balances with respect to income should be important even though this contention has to be proven empirically. The second result concerns the specific shape of the demand curve for money with respect to the interest rate as it is illustrated by the solid line in figure 2. Its sharp decline arises from two sources: from the decreasing marginal substitution ratio $\partial N^P/\partial m$ being equal to aT/m^2 according to relation (5a), and from the decreasing marginal productivity of labor $\partial y/\partial N^P$ due to the standard interpretation of the production function (1). One implication of this second result is straightforward for the long-run production effect of inflation. An economy will suffer from production losses at an increasing rate the higher the rate of inflation is. Another implication concerns the optimum quantity of money. The policy of converging the economy toward the optimal money supply will result in production gains, but at a decreasing rate, so that the "practical" optimality solution may be less than the "satiety" level of real cash balances if there is any.

References

Bailey, Martin J. *National Income and the Price Level.* 2d ed. New York: McGraw-Hill, 1971.

Brunner, Karl, and Meltzer, Allan. "The Uses of Money: Money in the Theory of an Exchange Economy." *A.E.R.* 61 (December 1971): 784–805.

Mundell, Robert A. *Monetary Theory.* Pacific Palisades, Calif.: Goodyear, 1971.

Sinai, Allen, and Stokes, Houston H. "Real Money Balances: An Omitted Variable from the Production Function." *Rev. Econ. and Statis.* 54 (August 1972): 290–96.

Book Reviews

G. D. H. Cole: An Intellectual Biography. By L. P. CARPENTER.
New York: Cambridge University Press, 1973. Pp. 271. $14.95.

G. D. H. Cole was equipped with an extreme literary facility, indeed an almost compulsive fluency, and from about 1913, with his first important work *The World of Labour*, until 1960, with the fifth and final volume of his *History of Socialist Thought*, he turned out literally scores of books and pamphlets (Mr. Carpenter doesn't tell us precisely how many). Cole's writings ranged over social and economic history and biography, labor and trade-union organization, political and social theory and ideas, industrial organization, economic planning, money, and international affairs, in addition to detective stories and several volumes of verse (he is represented in *The Oxford Book of English Twentieth Century Verse*).

Apparently quite deliberately, Cole was prepared to sacrifice depth and quality to achieve breadth and effect in his work as a whole. Carpenter quotes three times over (pp. 3, 10, 252) a passage Cole wrote about William Morris, but which, with reference to Cole himself, provides (according to Carpenter) "the definitive answer to those who rebuked him for having written too much: 'Though a man's work may fall short of greatness if he attempts too many things, it does not at all follow that he would do better in attempting less. For the truth may be that he wants to do and say so much that he is much more concerned to get it done and said than to do one thing, or a few things, supremely well.' "

However valid this argument may be in general terms, or whatever applicability it may have regarding Cole himself, what should "an intellectual biography" of such a man contain? How should it be constructed, what kinds of explanations or interpretations should be sought, and how should they be fitted together?

One line of approach would be in terms of some kind of rather deep psychological, or "psycho"-, analysis. Carpenter broaches this on page 1 and abandons it on page 2 after a reference to Erikson and "Freudian developmental models." As so often happens, the kind of evidence required is simply not available, and this was sufficient to deter Carpenter in the original construction of his thesis.

Another kind of "intellectual biography" might have been composed mainly in terms of Cole's contributions to the "internal" problems and development of his subject. Such an approach might have proved adequate if Cole's subject had been, say, one like pure mathematics. But Cole's "subject," or range of subjects, was vast and sprawling, interacting vitally with the economic, political, and social events of his time. So what Carpenter gives us is, for the most part, a discussion of Cole's writings, in broadly chronological order, with a background of contemporary history rather intermittently sketched in. It is, however, hardly possible in 250 pages to get to grips sufficiently firmly with the vast economic, political, and social background of Cole's writings to illuminate satisfactorily some of the main questions thrown up; for example, that of Cole's transition from the guild socialism of his youth to his later inclinations toward Fabianism and economic planning.

Carpenter has chosen a rich subject of continuing interest and importance and has energetically pursued with his questions a number of Cole's friends and associates, mostly in Oxford. But it is unfortunate that Mrs. Cole's life of her husband seems to have appeared just too late for Carpenter to make any use of it. Moreover he doesn't formulate any very sharply pointed questions or reach any sharply pointed conclusions. It is reasonable for a biographer to be on the side of his subject. But surely he should recognize, formulate, and attempt to answer the more obvious questions and criticisms which arise, especially those concerning so extremely controversial a figure as Cole. To start with, some reconciliation or explanation might have been attempted regarding Cole's belief in socialism "as a complete alternative way of living" (p. 11) and his own narrowly English upper-middle-class tastes and life-style. More theoretically, Carpenter might have offered some critical comments on Cole's proclaiming himself a champion of "freedom," while at the same time proclaiming a devoted admiration—almost rivaling that of the Webbs, Sweezy, Dobb, and Joan Robinson—for the achievements of Stalin.

We are certainly much in need of a critical survey of socialist ideas in Britain since the beginning of this century and notably of those of the leading triumvirate of Tawney, Laski, and Cole. This book might be regarded as a contribution in this direction.

T. W. Hutchison

University of Birmingham

Development of the Monetary Sector: Prediction and Policy Analysis in the FRB-MIT-Penn Model. By J. Phillip Cooper.
Lexington, Mass.: Lexington Books, 1974.

Econometric models in recent years have been playing an important role in economic analysis, and numerous models exist today which are widely used by both business and government. One of the most prominent, the FRB-MIT-Penn (FMP), is the focal point of Cooper's book. The book contains four chapters, three of which are edited or slightly altered versions of previous works by Cooper and associates. Chapter 1 is the Modigliani, Rasche, and Cooper paper (1970), which describes the structure and estimation of the FMP's monetary sector. Chapter 2 is the Cooper and Nelson paper (1972), and chapters 2 and 3 describe some forecasting properties of the complete FMP model. Chapter 4 is the Cooper and Fischer papers (1972*a*, 1972*b*), which describe what the model implies about the effects of some alternative monetary rules.

Given the proliferation of econometric models today, it would be helpful to users if methods and procedures for evaluating these models were devised. This seems to be one of Cooper's major objectives. Others include an evaluation of the estimation techniques used in constructing the monetary sector of the FMP model, an attempt to resolve the fixed monetary growth-rate rule controversy, and an attempt to resolve controversy involving structural versus reduced-form models. None of these diverse objectives, however, is achieved.

The monetary sector developed in the first chapter consists of four basic equations: (1) a demand-deposit equation, (2) a free-reserve equation, (3) a reserve identity connecting the monetary instrument and unborrowed reserves to member bank deposits, and (4) an identity connecting member bank demand deposits to total demand deposits held by the public. The system determines four unknowns: free reserves, member bank deposits, total deposits, and the three-month Treasury bill rate. Equations (1) and (2) are estimated over the data period 1955: I–1966: IV, first using ordinary least squares (OLS) and then using a two-stage estimator with, among other things, unborrowed reserves as an instrument.

The estimators are then compared in the following way. For each estimator, the model is dynamically simulated from 1955: I–1968: I. It is found that within the data period, OLS produces a smaller root-mean-square error (RMSE) for total deposits but larger for the bill rate, while in the postsample period the two-stage estimator is unambiguously more accurate. The author concludes that the two-stage procedure provides "more reliable" estimates of the underlying structure (p. 70). To draw this conclusion it must be true in general that inconsistent estimates produce larger RMSEs than consistent estimates. But each element of such a RMSE is a residual from a constrained reduced-form equation, constrained by the overidentifying restrictions imposed by the structure, and thus is a complicated function of structural coefficients and residuals. There is no simple way to determine the effects of inconsistent estimates on RMSEs and, consequently, there is no simple way to interpret Cooper's results. More generally, there is no basis for drawing conclusions about different estimators from results on a single data set for which the true model is not known.

The second chapter compares the predictive performance of the complete FMP model to that of single-equation time series models which were selected from the general class of discrete, linear stochastic processes of integrated autoregressive moving average (ARIMA) form. The investigation is confined to one-quarter-ahead extrapolations within the sample period (1956: I–1966: IV) and over a postsample period (1967: I–1968:IV). Relying primarily on RMSE, Cooper finds that the FMP model is "more accurate" on four of six "key" variables. It is "more accurate" on nominal GNP, real GNP, and the long- and short-term interest rates, while "less accurate" on the GNP deflator and the unemployment rate.

What are we to conclude from this experiment? In this case, the FMP RMSEs examined are residual standard errors of the implied one-period reduced-form equations. Since the coefficients of those equations are constrained by the overidentifying restrictions imposed by the structure assumed for the model, there evidently exists a class of unconstrained reduced-form equations that have smaller RMSEs than those implied by the model. Had Cooper chosen such equations instead of his particular ARIMA equations, the FMP model would have been "less accurate" on every variable. Would that imply that overidentifying restrictions should never be imposed? Nor does it help much to make Cooper's comparisons for data outside the initial sample period. If there has been no structural change, the relatively unconstrained equations will continue to exhibit the lower RMSEs. If there has been structural change, one or both sets of reduced forms may fail; but a comparison of RMSEs across the sets is less informative than a comparison of within- and without-sample-period RMSEs for each set.

Cooper begins chapter 3 by claiming that we need many replications in order to have "meaningful" results (p. 121). But the hope that the author will eventually present a "meaningful" test is short lived. The simulations in this chapter differ from those of the previous chapter in that additive disturbances are set equal to pseudorandom numbers which have statistical properties consistent with estimated residual variances. Twenty-five replications of both dynamic and one-quarter simulations for the period 1956: I–1968: IV are produced. We are then provided with 20 figures and 10 tables which report various aspects of simulation results for five selected variables—real GNP, nominal GNP, GNP deflator, and long- and short-term interest rates. The figures contain time paths of the mean stochastic solutions, the stochastic mean plus and minus the standard deviation of the stochastic solutions, the minimum and maximum solutions, the deterministic solutions, and the actual data. The tables contain a number of measures of goodness of fit, including RMSE, mean absolute error, and mean error.

Cooper goes on to discuss some of the general tendencies found in these stochastic results. For example, in the dynamic simulations (1) the stochastic means are close to the deterministic solution, suggesting that the nonlinearity of the FMP model is of little consequence; (2) the standard deviation and the range of the stochastic solutions increase with time; and (3) the amplitudes of peaks and troughs tend to be underestimated. However, beyond a few of these general observations, no conclusions are drawn about the reliability of the FMP model. Cooper fails to recognize that a test of a model's reliability could be based on a comparison between the residuals outside the estimation period and the estimated within-period standard errors.

Cooper claims to use the stochastic data to judge the reliability and stability of so-called reduced-form models. Using actual data and data generated from both the deterministic and stochastic simulations of the FMP model with unborrowed reserves exogenous, he reestimates the Saint Louis "expenditure" equation. Comparing the coefficients derived from these data sources, he finds that the coefficients estimated on stochastic data are closer to zero and not as stable as those derived from the actual and deterministic data. Cooper, for reasons I find difficult to understand, regards this comparison as a test of the Saint Louis equation and more generally as a test of reduced-form models. At best, however, it seems to be a test of the FMP model and of the assumption that unborrowed reserves were indeed exogenous over the data period. His finding that the coefficients based on the stochastic data appear to be significantly different from those based on actual data contradicts this joint hypothesis.

There is also some confusion here about the meaning of exogeneity. On the basis of the chapter 1 model, in which the money supply is assumed to be endogenous and unborrowed reserves are the monetary instrument, Cooper criticizes the Saint Louis model for treating the money supply as an exogenous regressor (p. 165). Yet Cooper fails to test directly for exogeneity; and in fact, contrary to Cooper's supposition, Sims (1972) found the money supply to be statistically exogenous.

Chapter 4 examines rules for monetary policy, constant growth-rate rules, and more general feedback rules for the money supply. Deterministic and stochastic simulations of the FMP model are used to determine whether the variability (measured by RMSE and the standard deviation) of rates of inflation and unemployment produced under a constant growth-rate rule can be reduced by following some simple feedback rule. In these simulations all exogenous variables other than the money supply are set at their historical levels. The money supply is determined by the chosen rule plus a constraint on the bill rate. In the stochastic simulations the bill rate is constrained to be greater than 0.5 percent. This is accomplished by setting the money supply at a level consistent with 0.5 percent whenever a rule would have produced a lower rate. Since the problem lies with the FMP model's linear money demand equation, which does not contain a liquidity trap, even at zero, the appropriate constraint should have been imposed directly on the demand curve. Thus, whenever a rule would have produced a rate below 0.5 percent, instead of shifting the money supply, the bill rate should have been set at 0.5. Not having done this, the rules actually examined are probably much different from those described.

In any case, such an experiment contains little, if any, information pertinent to the rules controversy. Given a large nonlinear model like the FMP model in which expectations of future variables are generated by the standard "adaptive" mechanism, the optimum rule is likely to be a complex feedback rule. So once we accept the model, there is little controversy. Cooper alludes to this issue when he states that a potential difficulty with this type of study is the possibility that adoption of a rule per se might modify behavioral relations of the economy in such a way as to reduce uncertainty and increase the stability of the economy (pp. 193–94). The problem, however, is more general than he suggests. It is not just that the introduction of rules will change behavior, but each new rule may give rise to new behavioral relationships (Lucas 1973). Moreover, if expectations of future variables are formed "rationally," one rule may be as good as another (Sargent 1973).

In summary, Cooper promises his reader much more than he delivers. For those unfamiliar with Cooper's original works, this book may provide some useful information, but for most it will be a review of outdated results. Their hopes of finding procedures and tests for evaluating econometric models, of finding a specific test of the FMP model and its estimation techniques, and of finding new results on the rules controversy will be disappointed.

ARTHUR J. ROLNICK

Federal Reserve Bank of Minneapolis

References

Cooper, J. Phillip, and Fischer, Stanley. "Simulation of Monetary Rules in the FRB-MIT-Penn Model." *J. Money, Credit, and Banking* (May 1972), pp. 384–96. (*a*)

———. "Stochastic Simulation of Monetary Rules in Two Macroeconometric Models." *J. American Statis. Assoc.* (December 1972), pp. 750–60. (*b*)

Cooper, J. Phillip, and Nelson, Charles R. "The Ex Ante Prediction Performance of the St. Louis and FRB-MIT-Penn Econometric Model." CMSBE Report 7236. Mimeographed. University of Chicago, 1972.

Lucas, Robert E. "Econometric Policy Evaluation: A Critique." Manuscript, Carnegie-Mellon University, April 1973.

Modigliani, Franco; Rasche, Robert H.; and Cooper, J. Phillip. "Central Bank Policy, the Money Supply and the Short-Term Rate of Interest." *J. Money, Credit, and Banking* (May 1970), pp. 166–218.

Sargent, Thomas J. "Rational Expectations, the Real Rate of Interest, and the Natural Rate of Unemployment." *Brookings Papers Econ. Activity*, no. 2 (1973), pp. 480–92.

Sims, Christopher A. "Money, Income and Causality." *A.E.R.* (September 1972), pp. 540–52.

Unemployment and Inflation: The New Economics of the Wage-Price Spiral. By ROGER L. MILLER and RABURN M. WILLIAMS.
Saint Paul, Minn.: West Publishing Co., 1974.

Very little in the book *Unemployment and Inflation*, by Roger Miller and Raburn Williams, is original research, as the authors concede, so that the work must be judged by its ability to synthesize and explain current strands of macroeconomic research. On the whole, the macroeconomic analysis which covers chapters 1–5 is very weak, while the microeconomics of employment presented in chapters 6 and 7, although not original, is very clearly and interestingly presented. This contrast is unfortunate, since the authors wish the text to be primarily of macroeconomic interest, and indeed the profession would welcome a work synthesizing the new "expectational" approach to aggregate analysis.

Chapter 2, the analysis of inflation, is certainly the weakest chapter in the book. The presentation is extremely choppy, with one disconnected model following another, so that when the chapter is over, the reader realizes that nothing more than the standard $IS = LM$ has been presented. I believe that this chapter should have been used by the authors to stress the role of money in the determination of the absolute price level and to expose the many "fallacies of composition" that mark much public discussion of inflation.

Chapter 3, a theory of investment in unemployment, is essentially a microeconomic chapter. The concept of the optimal search and stopping time has already been applied to the case of labor by McCall, Gronau, Hynes, and others; I think the authors should acknowledge these more fully. I believe that the authors should also realize that their expression for the optimal reservation wage, W_0 (eqq. [3]–[6]), is not an explicit solution, and hence cannot be manipulated in the manner that they do throughout most of the chapter. This chapter should also have been used to relate search unemployment to other forms discussed in the literature as queue, informational, structural, rationed, etc. One does not know whether the authors believe that search unemployment is all that is relevant for the discussion of aggregate-cycle theory.

Chapter 4 begins with a very incomplete discussion of the trade-off between inflation and unemployment. There is virtually no analysis of the welfare losses of either anticipated or unanticipated inflation which has an extensive literature. And even more important, since the book is more a study of unemployment, there is a total lack of treatment of the welfare aspects of unemployment. If unemployment is due to forecasting errors, as the authors suggest, then any unemployment rate differing from the natural rate, whether from above *or below*, must be an inferior welfare equilibrium. This, of course, leads to the question of the Pareto optimality of the natural rate itself, which is of crucial importance and to which the authors give no attention. Are there externalities or nonconvexities in the labor market which might call for public policy to change the natural rate? If unemployment is caused by forecast errors, then is there a Pareto-optimal amount of information produced in the market? Is this information being efficiently utilized? Is there a welfare gain through higher demand just because bid-ask spreads for labor and commodities are reduced? (This has been suggested by Phelps.) These appear to me to be critical questions, for if unemployment is really only caused by fooling the public, society is worse off, even in the short run, by reducing unemployment.

The relationship between unemployment and price expectations is, of course, the crux of the modern theory of the Phillips curve. Miller and Williams demonstrate the well-known clockwise loops of inflation and unemployment which prevail under the adaptive-expectations mechanism so widely used. It is interesting that in the original Phillips analysis of British data, counterclockwise loops were formed, but this is not mentioned by the authors. Miller and Williams correctly indicate that only if prices follow a particular stochastic process (namely, white noise superimposed on a random walk, or an $IMA[1, 1]$ process in time-series jargon) will adaptive expectations be optimal or "rational." Therefore, a reader will readily ask, Do prices follow such a process? Would one theoretically expect them to follow this process? If prices are determined essentially by the money stock plus random shocks, an $IMA[1, 1]$ process for prices would have implications for the stochastic process determining high-powered money creation. What do the data show? It is unfortunate that the authors totally ignore these critical questions. The crux of the current rational expectation "onslaught"—and the authors are fully aware of its importance—is that systematic, expected components of money creation "wash out," that is, do not change (on first approximation), the real variables of the system, even in the short run. The autoregressive characteristics of the remaining random components become critical for optimal forecasting, and the whole shape of the Phillips curve is dependent on these forecasts.

Another point that the authors miss is that forecast errors must be serially uncorrelated. But if deviations of the unemployment rate from the natural rate are due to forecast errors only, then these deviations must be uncorrelated. But we know that the unemployment rate is autocorrelated, so we must assume that the natural rate is. How can the Great Depression, with unemployment rates above 15 percent quarter after quarter, be explained? According to rational expectations, it can only be explained by a sharply higher natural rate. Certainly Friedman and others believe that the natural rate was much higher during the Depression, primarily due to government intervention. But if most of the fluctuations of the observed unemployment rate are due to shifting of the natural rate, especially for large movements such as wars and depressions, the rational-expectations theory, although correct, seems unimportant.

Conspicuously lacking from the book is a discussion of indexing and escalated contracts. There are dozens of important questions that arise with indexing. Do incorrect price forecasts make any difference in an "indexed world"? If unemployment is caused by current forecast errors, might even the business cycle be eliminated or sharply reduced by indexing? If indexing is so favorable, why have creditors and debtors (and the government) been so reluctant to adopt such contracts? Theoretical work seems far behind events on this topic, and it is disappointing that the authors could not have foreseen its importance.

As I mentioned earlier, chapters 6 and 7 are an interesting microeconomic presentation of the demand for labor and how it is affected by the policy, demand, and supply parameters of the system. However, little macroeconomic synthesis is attempted, and I find this particularly disappointing. The entire analysis of how policy affects the *absolute* price level should be done in terms of the effect on the real variables appearing in the demand function for real money balances. For instance, one can analyze the "inflationary" effect of a higher minimum wage in the following manner. For a given nominal money supply, real employment would be constricted (assuming a fairly competitive labor sector), reducing real output. The rate of interest might decrease due to the increase in the capital-labor ratio, at least in the short run. The decreased income works in the direction of a higher equilibrium price level, while the decreased interest rate works toward a lower level due to the increased demand for real balances. But there are other factors. Some laborers are involuntarily consuming leisure while others have a higher wage or opportunity cost of leisure. One would hence have to examine the degree of complementarity of transactions services (as a proxy for real balances) and leisure. This is how I believe microeconomic policy measures should be studied for their price level effects, but Miller and Williams do not proceed along these lines.

The final chapter, "Policy-making in the Post-New Economic Era," although only two pages long, is an excellent summary of the rational-expectations approach to macroeconomics. In fact, an interesting interpretation of the *cost* of making predictions (assumed zero and an obvious weakness of the literature so far) is even used to interpret prices and unemployment in the 1960s. The authors do

have a remarkable clarity of style—an ability to explain complicated processes in simple terms. It is too bad that more innovative interpretations had not marked some of the earlier chapters. This would have given the work greater cohesiveness and would have ameliorated the disjointed topicality of most of the macroeconomic discussion.

JEREMY J. SIEGEL

University of Chicago

Schooling, Experience, and Earnings. By JACOB MINCER.
New York: National Bureau of Economic Research, 1974.

In this compact volume, Jacob Mincer provides a theory and extensive empirical evidence about the process of the accumulation of human capital on the job. The main features of the theory will be familiar to readers of Mincer's earlier work, but the empirical evidence is new. The book is uncompromisingly scientific rather than polemical—it is a detailed presentation of findings within a particular intellectual framework, not an argument for that framework against alternatives. Most of the testing of the theory takes the form of checks for internal consistency within the established framework. The book is aimed at the believer, not the skeptic, though the skeptic is bound to be impressed by Mincer's careful study of the data and the substantial accord of his findings with the predictions of the theory.

Mincer's basic theoretical contribution is the working out of the implications of investment on the job in skills and general experience. He views the employer as a purchaser of labor services and a seller of training at the same time. The observed wage is the return to the worker's services, including the return to his earlier investments in training, less the value of his current purchases of training. Comparisons of observed wages among workers gives a mistaken view of the differences in their true incomes by overlooking the values of investments that are deducted from gross wages. Both the current flow and the past stock of these investments are unobservable, however, so the theory at this stage amounts to nothing more than a relabeling of unexplained differences among workers as differences in unobserved stocks and flows of human capital acquired on the job. The power of Mincer's theory arises from his hypothesis that workers invest in human capital so as to maximize the present value of lifetime earnings. The optimal pattern of investment starts at a high level and declines over the lifetime. The gross wage starts at a low level representing the return to raw labor and formal education and rises throughout the lifetime as capital is accumulated on the job. The observed net wage rises more rapidly because it is the difference between the growing return to labor and the shrinking flow of investments. A central feature of the theory is its prediction that the net wage will equal the gross return to inexperienced labor after a certain period in the labor force and that this "overtaking point" depends only on the discount rate. The wages of workers at the overtaking point can be explained by observable variables alone, principally formal education. The group of workers at the overtaking point (around 8 years of experience) is uniquely important in Mincer's sample.

Mincer's theoretical exposition represents the polar extreme to the anecdotes and institutional detail typical of most writing in labor economics. In his view, the theory of training and wages is an application of the standard economic theory of optimal choice subject to a budget constraint. In particular, he does nothing to convince the reader that workers face an operational choice about their purchases of training on the job. Critics who believe that institutional rigidities render the economic theory of labor supply meaningless will argue along the same lines that workers are forced to accept a predetermined amount of training. The central argument of good economics in general and the theory of human capital in particular that economic forces determine institutions and not the other way around is taken wholly for granted in this book and limits its readership unnecessarily, in my view. Even within the theory of human capital, Mincer deals only briefly with Gary Becker's important distinction between general and specific human capital. Workers cannot rely on capturing the return to investment in skills uniquely applicable to a particular employer and so will not follow Mincer's optimal path in acquiring them. Mincer is almost certainly right that the returns to general human capital dominate the returns to specific capital for the typical worker, but the point deserves more discussion in the book.

Mincer's empirical work encounters problems in measuring wages that are familiar to all workers in this area. The Bureau of the Census collects data on annual earnings and annual weeks of work for the previous year (1959 in this case) and hours of work in the previous week. Many investigators have measured hourly wages as annual earnings divided by the product of annual weeks and weekly hours, a procedure known to have some biases. Mincer instead uses annual earnings itself as his dependent variable. In some cases he treats annual earnings as if it were proportional to the wage (implicitly assuming that annual hours of work do not vary systematically in the population), and in other cases he includes annual weeks on the right-hand side. The second procedure seems especially questionable, as his estimated elasticity of earnings with respect to weeks is not 1.0 but about 1.2. The theory of labor supply suggests that weeks of work is an endogenous variable, as Mincer concedes. This complicates the empirical verification of the major hypothesis of the paper that a large fraction of the dispersion of wages in the population can be explained by the observed determinants of wages. In the subset of the data for workers at the overtaking point, only 31 percent of the variance in annual earnings is explained by the observed determinant, years of education; but when weeks of work is added, the fraction rises to 58 percent. It would be interesting to have the explained fraction of the variance of a wage variable of some kind, perhaps even annual earnings divided by annual weeks.

Schooling, Experience, and Earnings deserves comparison with Christopher Jencks's influential recent book, *Inequality: A Reassessment of the Effect of Family and Schooling* (New York: Basic, 1972). Jencks's writing is frankly polemical, arguing that redistributing the determinants of income does almost nothing to redistribute income itself. His evidence is mainly the low explanatory power of equations relating income to its observed determinants. In his work, education alone explains only 11 percent of the variance of income, and even when measures of family background, test scores, and occupational status are added, the fraction rises only to 22 percent. In sharp contrast, Mincer claims that "...schooling and post-school investment accounted for close to two-thirds of the inequality of earnings of adult, white, urban men in the United States in 1959" (p. 96). The two studies use different bodies of data but cover essentially the same population. They can be reconciled roughly as follows: First, Jencks studies income, which has more dispersion than earnings and is harder to explain. Second, Jencks uses wide age categories and attributes little explanatory power to age or experience. This is probably the single largest discrepancy—Mincer reports that education alone explains only 7 percent of the variation in earnings in his sample, against 31 percent when education and experience are both included. Even on Jencks's own ground, Mincer demonstrates that more careful scientific work weakens the conclusion that the observed determinants of income have little influence on its dispersion. Mincer then goes on to his figure of two-thirds by two less firmly supported inferences. In the overtaking group, where earnings are uncontaminated by investments in training or the return to past training, the residual variance of the log of earnings is 33 percent. In an equation where the contribution to earnings of investment and return was observed, the residual variance should be about the same. The total variance of log earnings is 0.67, so the fraction of variance explained by the hypothetical equation is just under one-half. Mincer believes that better measures of education could raise this to 55 percent. Finally, equations that take credit for the explanatory power of annual weeks of work explain 69–78 percent of the variance, depending on certain assumptions. Mincer's two-thirds is a rough average of the three figures, 55, 69, 78. At least from the point of view of Jencks's proposition about the effect of redistributing the determinants of income, 55 percent seems the highest defensible claim. Weeks of work are jointly determined with earnings and cannot reasonably be labeled as determinants. Even so, 55 percent is far above 22 percent, and Mincer's evidence could form the basis of a strong counterattack on Jencks's position by those who believe that earnings depend closely on rational choices about investments in education and training. Such a counterattack would be very much out of the spirit of Mincer's book, however.

This book makes a substantial contribution to scientific knowledge about the earnings of individuals and the distribution of earnings in the population. It should open up fruitful new lines of research on the centrally important questions of income distribution and human capital.

Robert E. Hall

Massachusetts Institute of Technology

Distribution Effects and the Aggregate Consumption Function

Alan S. Blinder

Princeton University

This paper investigates whether and how the distribution of income affects the fraction of disposable income which is consumed. It is shown that a slight generalization of the permanent income or "life-cycle" model of consumption makes each individual's lifetime marginal propensity to consume a fraction of his lifetime disposable resources unless two taste parameters are equal. In considering what this implies for the aggregate consumption function, the tenuous connection between theoretical constructs and observed facts is stressed. Previous empirical work on the subject is criticized for failing to define properly either "income distribution" or "consumption," and a new test, based on the theory, is outlined. Because of data limitations, a number of compromises with this ideal test must be made, and several alternative models are estimated. On the whole, they suggest that equalizing the distribution of income would either leave aggregate consumption unchanged or diminish it slightly.

I. Introduction

Does the manner in which a given amount of income is distributed affect the fraction of it that is consumed? In the early post-Keynesian days it was commonly assumed, presumably on the basis of Keynes's own intuition, that it does—in particular that equalization of the income distribution would increase consumption. With the publication of Kuznets's (1942) and Goldsmith's (1955) data, and the ascendancy of the Friedman (1957) and Modigliani and Brumberg (1954) models of consumer

---

I would like to thank, without implicating, Robert Barro, William Branson, Michael Darby, Ray Fair, Malcolm Fisher, Milton Friedman, Stephen Goldfeld, Michael Hurd, James MacKinnon, and Michael Rothschild for helpful comments which materially improved the content of this paper. The research reported herein has been generously supported by the National Science Foundation.

[*Journal of Political Economy*, 1975, vol. 83, no. 3]
© 1975 by The University of Chicago. All rights reserved.

behavior, this view fell into disrepute in academic circles. It was supplanted by the view that marginal, and perhaps even average, propensities to consume are constant over the income distribution. Of course, this "modern" view does not accord very well with intuition. It does, after all, seem "obvious" to most people, especially those not schooled in macroeconomics, that the rich save proportionately more than the poor, even at the margin.

While what is "obvious" is not always true, I was somewhat shocked to discover that the notion that aggregate consumption is independent of the income distribution has never been subjected to a direct empirical test. That is, the hypothesis that the *size distribution* of income does not affect consumption has never been treated as a special case of a more general class of consumption function and tested by standard statistical techniques. I propose to do so in this paper.

Let me begin with a confession. At the outset of this research I hoped to establish:

> *PROPOSITION A:* The marginal propensity to consume of an individual falls as his disposable income rises.

And therefore:

> *PROPOSITION B:* Out of any given total disposable income, a larger share is spent on consumption when income is more equally distributed.

As it turns out, while both theory and empirical evidence lend at least some support to A, they do not support B. In fact, as I explain shortly, Proposition B does not follow from Proposition A. What does follow is the similar-sounding proposition:

> *PROPOSITION C:* If income is taken from one individual and given to another individual who is identical in all relevant respects save that his income is higher, then total consumption will decline.

The next section develops the theoretical equipment necessary to investigate the effects of redistribution on aggregate consumption. I derive a plausible condition on individual utility functions which is sufficient to guarantee Proposition A, and then establish (as should be obvious) that A implies C. Section III explains why I view previous tests of the effect of income inequality on aggregate consumption as inconclusive, and derives a model for testing Proposition B in the context of the permanent income theory. Section IV explains some compromises that had to be made because of weaknesses in the data, and presents the empirical results I have obtained with a compromise model. These results suggest that a rise in income inequality, disposable income held constant, would either have no effect on consumption or would actually *increase* it. That is to say, while Propositions A and C may be true, the obverse of B is given at least mild support by postwar American data. Section V offers a variety of possible explanations for this result, and the last section is a brief summary.

II. The Implications of Pure Theory

A. Optimal Life-Cycle Consumption

By now the derivation of the aggregate consumption function from a Fisherian model of intertemporal utility maximization, as pioneered by Modigliani and Brumberg (1954) and Friedman (1957), has achieved widespread acceptance. In this theory, the consumer chooses the time path for consumption, $c(t)$, which maximizes lifetime utility, subject to the constraint that the present discounted value of all consumption, plus the present discounted value of the bequest (if any), is equal to lifetime disposable resources, W. That is,

$$\int_0^T c(t)e^{-rt} + K_T e^{-rT} = W, \tag{1}$$

where t is age, r is the rate of interest, T is the length of life, K_T is the bequest, and W is defined as the sum of the inheritance plus the present discounted value of earned income.[1] In order to get the typical "life-cycle" or "permanent income" result that consumption at each instant is proportional to W, the lifetime utility functional must be of the form

$$U = \int_0^T \frac{c(t)^{1-\delta}}{1-\delta} e^{-\rho t}\, dt + \frac{bK_T^{1-\beta}}{1-\beta} : \quad \delta, \quad \beta > 0; \quad b \geq 0, \tag{2}$$

with the further stipulation that $\delta = \beta$.[2] Thus (2) represents a minor generalization of the Modigliani-Brumberg-Friedman (henceforth MBF) model—a generalization with important consequences for the question of distribution effects.

The maximization of (2) subject to (1) is a well-known problem, which was first studied by Strotz (1955–56). A heuristic method of solution is given in the Appendix. For present purposes, it suffices to note that the optimal plan is given implicitly by the following equations:

---

[1] Equation (1) is the budget constraint only under the assumption of a perfect capital market. This same assumption allows me to collapse all gifts to heirs into a single number, K_T. That is, an *inter vivos* gift of G given at age t is *equivalent*—for both donor and recipient—to a gift of $Ge^{r(T-t)}$ at death.

[2] This is implied by some results of Yaari (1964) and is developed in some detail in Blinder (1974, Chap. 2).

$$c(t) = c_0 e^{gt}, \qquad \text{where } g \equiv (r - \rho)/\delta; \tag{3}$$

$$c_0 = \phi(r, \rho, \delta, T)(W - K_T e^{-rT}); \tag{4}$$

$$K_T = (be^{rT})^{1/\beta} c_0^{\delta/\beta}, \tag{5}$$

where $\phi(\cdot)$ *is a known function specified in the Appendix.*

The strict MBF model holds that consumption at each instant is proportional to W, with the constant of proportionality dependent on age (t), the length of life (T), the rate of interest (r), and tastes. This result follows from (3)–(5) under two sets of circumstances. The first is $b = 0$. This is the strict life-cycle model of Modigliani-Brumberg-Ando (MBA). If there is no utility from bequests, then the optimal K_T will be zero for every person, as is clear from (5). By (3) and (4), then, c_0 (and hence all $c[t]$) will be proportional to W. Specifically, $c(t) = \phi e^{gt} W$. The second condition is $\delta = \beta$. This is a modification mentioned by Modigliani and Ando (1960). In this case K_T is proportional to c_0 by (5), so that (3)–(5) can be solved to yield

$$c(t) = \left(\frac{\phi e^{gt}}{1 + \phi b^{1/\delta} e^{rT(1-\delta)/\delta}} \right) \cdot W.$$

It is important to note that these are the *only* two cases which give rise to strict proportionality, that is, a constant lifetime marginal propensity to consume (MPC). Extending the life-cycle model to allow for the bequest motive destroys the proportionality property unless $\delta = \beta$. The lifetime MPC in the more general case can be found by implicit differentiation in (3)–(5). The answer turns out to be

$$\frac{\partial c^*}{\partial W} = \left(1 + \frac{\delta}{\beta} b^{1/\beta} e^{rT(1-\beta)/\beta} \phi^{\delta/\beta} c^{*(\delta/\beta)-1} \right)^{-1},$$

where

$$c^* \equiv \int_0^T c(t) e^{-rt} \, dt$$

is lifetime consumption. In words, the lifetime MPC is smaller than unity as long as $b > 0$, and is decreasing with W if $\delta > \beta$ or increasing with W if $\beta > \delta$.

What do these conditions mean? If β, the elasticity of the marginal utility of bequests, exceeds δ, the elasticity of the marginal utility of consumption, then consumption is the luxury good, that is, has a wealth elasticity greater than unity. Conversely, if $\delta > \beta$, then bequests are the luxury good. It seems plausible, to me at least, that bequests should be the luxury good;[3] but, in the absence of the requisite empirical evidence, each reader is free to make his own judgment. My only purpose here is to establish that it is possible, within the basic MBF model, to have an MPC which either rises or falls with income.

B. The Effect of Redistribution on Aggregate Consumption

I shall now prove that, if the MPC declines with W, an increase in income inequality must reduce consumption. Conversely, if the MPC actually rises with W, a rise in inequality will increase consumption.

Consider a population of individuals identical in every respect save permanent income. Let y denote permanent income, defined as the flow equivalent of the stock of lifetime resources, so that y is proportional to W. Then the model of consumption behavior just developed implies $c = c(y)$, $1 > c'(y) > 0$, and $c''(y) \gtreqless 0$, according as $\delta \lesseqgtr \beta$. Let the distribution of permanent income be given by a density function $f(y, d)$, where d is a very general indicator of inequality to be explained shortly; and let $F(y, d)$ be the corresponding cumulative distribution function. Finally, let μ, a, and b denote, respectively, the average, lowest, and highest permanent income in the population.

It is easily established that[4]

$$\mu = b - \int_a^b F(y, d)\, dy. \tag{6}$$

The parameter d represents what Rothschild and Stiglitz (1970) have termed a "mean preserving spread." That is, a rise in d signifies a sequence of transfers from poorer persons to richer ones (called "regressive transfers")

---

[3] Becker's recent analysis of intergenerational transfers (1974) provides some theoretical support for the notion that $\delta > \beta$. Consider, for simplicity, a two-generation family. Letting c_i denote lifetime consumption, W_i denote lifetime wealth, and m_i denote the part of W_i not inherited, the budget constraints for generations $i = 1, 2$ would be $c_1 + k = W_1$ and $c_2 = W_2 \equiv m_2 + k$, where k is the bequest from generation 1 to generation 2, and all quantities are discounted to a common date ($c_2 = W_2$, since generation 2 leaves no bequest). Becker observes that the two budget constraints can be collapsed to $c_1 + c_2 = W_1 + m_2 \equiv S$, where S is what he calls "social income." He then shows that if the elasticity of c_1 (and therefore also of c_2) with respect to S is approximately unity, the elasticity of k with respect to W_1 (which is my δ/β) must exceed unity. The proof is almost immediate. Let η_{xy} denote the elasticity of x with respect to y. The budget constraint for generation 1 implies that a weighted average of $\eta_{c_1W_1}$ and η_{kW_1} equals unity. But $\eta_{c_1S} = 1$ implies that $\eta_{c_1W_1} = (W_1/S)\eta_{c_1S} < 1$. It is thus clear that $\eta_{kW_1} > 1$. QED.

[4] Proof: By definition

$$\mu = \int_a^b yf(y, d)\, dy.$$

Integrating this by parts yields

$$\mu = yF(y, d)\Big|_a^b - \int_a^b F(y, d)\, dy.$$

Equation (6) follows by noting that $F(b, d) = 1$, $F(a, d) = 0$.

which leave the mean unchanged. I add the further stipulation (solely for convenience) that the maximum and minimum incomes are also unaffected, so that a change in d must satisfy the following:

$$\frac{\partial a}{\partial d} = \frac{\partial b}{\partial d} = 0; \tag{7a}$$

$$F_d(y, d) \text{ is continuous on the interval } a \leq y \leq b; \tag{7b}$$

there is some y^* in the interval (a, b) such that

$$\begin{aligned} F_d(y, d) &\geq 0 \quad \text{for} \quad a \leq y \leq y^*, \quad \text{and} \\ F_d(y, d) &\leq 0 \quad \text{for} \quad y^* \leq y \leq b; \end{aligned} \tag{7c}$$

$$\frac{\partial \mu}{\partial d} = -\int_a^b F_d(y, d)\, dy = 0. \tag{7d}$$

The last requirement, that shifts in d leave the mean unchanged, follows from (6).

With the preliminaries thus established, the proof is quite direct.[5] Aggregate consumption is defined as

$$C = \int_a^b c(y) f(y, d)\, dy$$

so that the effect of an increase in inequality on aggregate consumption is

$$\frac{\partial C}{\partial d} = \int_a^b c(y) f_d(y, d)\, dy.$$

Integrating this by parts yields[6]

$$\frac{\partial C}{\partial d} = -\int_a^b c'(y) F_d(y, d)\, dy. \tag{8}$$

First consider (8) in what I take to be the more plausible of the two cases—the case where $c'(y)$ is falling. By (7), F_d is a continuous function which is positive when y is "low," is negative when y is "high," and integrates to zero over its entire range. The integral in (8) attaches higher weights to the (positive) values of F_d which occur when y is low than it does to the (negative) values of F_d which occur when y is high. Thus the integral must be positive, and $\partial C/\partial d$ must be negative. Conversely, if $c'(y)$ were a rising function of y, $\partial C/\partial d$ would be positive. Of course, if $c'(y)$ is constant, (7d) immediately implies that $\partial C/\partial d = 0$. In words, I have established:

> *PROPOSITION D:* A mean-preserving spread in the income distribution will decrease, leave unchanged, or increase aggre-

---

[5] The proof follows a suggestion by Rothschild and Stiglitz (1970, p. 237 n.).

[6] Again I use the facts that $F(b, d) = 1$, $F(a, d) = 0$.

gate consumption according as δ is greater than, equal to, or less than β.

Of course, Propositions A and C follow from D only with the added assumption that $\delta > \beta$.

It is worth pausing at this juncture to consider what has *not* been proven. Proposition D refers only to transfers within a group which is identical in all relevant respects save income. It is not applicable to transfers from one socioeconomic group to another (e.g., whites to blacks; men to women) if there is any reason to believe that tastes may differ in the two groups. Nor is it applicable to transfers where the age distribution of the donors differs from the age distribution of the recipients. Finally, the fact that the population consists of many age cohorts poses still another problem. Suppose $\delta > \beta$, and there is a decline in inequality in the older (donor) cohorts. As just noted, this would lead to greater average consumption within these cohorts. However, it would lead to lower average bequests, and hence to reduced consumption within the younger (recipient) cohorts. The macro consumption function will reflect both these changes.[7]

The practical implication of all this, of course, is that Proposition D gives no basis for predicting the effect on aggregate consumption of most real-world redistributions.[8] That is, it certainly does *not* establish Proposition B.

III. Testing for Distribution Effects

A. The Definition of Income Distribution

In the typical test for distribution effects in the aggregate consumption function, the income variable in the model is disaggregated into two or more components, and the hypothesis that the two (or more) regression coefficients are equal is tested. Suppose, for example, that the maintained hypothesis is represented by the estimating equation

$$C_t = aY_t + bC_{t-1} + u_t, \tag{9}$$

where Y_t is current disposable income (real or nominal; total or per capita) and C_t is consumer expenditures (defined symmetrically). A typical test is to divide Y_t into labor income (L_t) and property income (P_t), reformulate the model as

$$C_t = a_1L_t + a_2P_t + bC_{t-1} + u_t,$$

---

[7] I owe this last point to Robert Barro.

[8] The theory does, however, have testable implications that are not given in Proposition D. For example, *ceteris paribus*, a transfer from the young to the old will increase consumption (assuming g is positive).

TABLE 1

U.S. INCOME DISTRIBUTION IN 1962 BY COMPONENTS

DECILE GROUP	SHARE (%) IN: Wages and Salaries	Business and Property Income	Pensions and Annuities	Other Income
Lowest	0	0	11	4
Second	1	2	19	16
Third	3	4	16	24
Fourth	5	4	17	13
Fifth	8	6	4	9
Sixth	10	6	9	4
Seventh	13	6	5	7
Eighth	15	9	6	9
Ninth	18	13	4	9
Highest	27	51	8	5

SOURCE.—Projector, Weiss, and Thoresen (1969, table 4, p. 128).

and test $a_1 = a_2$ against the alternative $a_1 > a_2$. Generally the null hypothesis of no distribution effects cannot be rejected.[9]

What is the rationale for this test? The theory outlined in Section II suggests that MPCs might differ by income bracket, not by source of income. Presumably there is no reason for an individual to spend a different fraction of the marginal dollar, depending on whether it accrues in the form of wages or dividends. Suits (1963), in a review of the early literature on this question, suggested one possible justification: "... since functional shares vary by income bracket, taking account of functional distribution makes some allowance for the curvature in the consumption function...." That is, distributive shares might be a proxy for the distribution of income by size. This assumes, for example, that an increase in labor's share is reliably associated with an equalization in the size distribution.

How accurate is this assumption? The reader familiar with American income distribution statistics since World War II will be immediately suspicious, since labor's share has steadily increased while most conventional measures of inequality in the size distribution have either been constant or exhibited some upward drift. In point of fact, the division of national income between labor and capital has only a tenuous relation to the size distribution. Table 1 presents the distribution of four components of total income in the United States in 1962. Except in the highest decile the distributions of wages and salaries and of business and property income are not radically different in the sense that knowing whether a given dollar went to "labor" or to "capital" conveys relatively little information about where that dollar landed in the size distribution. In fact, it is not unambiguously clear that "nonlabor income" is distributed more unequally than "labor income," for the Lorenz curves cross. If pensions and annuities are grouped with business and property income, the resemblance is even stronger. Thus, testing whether aggregate consumption is sensitive to the *factor share* distribution is not a fair test of whether aggregate consumption is sensitive to the *size* distribution. This is the first error I set out to correct.

---

[9] The most recent example of this is Taylor (1971), who cannot reject $a_1 = a_2$ but does find significantly different MPCs out of transfers and other types of income.

B. *The Definition of Consumption*

The second error is the utilization of a theory of consumption to explain the behavior of *consumer expenditures.*[10] It is quite conceivable that the marginal propensity to *consume* could be falling with rising income while the marginal propensity to *spend on consumer goods and services* could be constant, or conversely.

To distinguish the various concepts of consumption, I introduce the following notation: C = consumption; CE = consumer expenditures, as defined in the national income accounts; CD = expenditures on consumer durables; and UD = use value of the stock of consumer durables, defined as the sum of depreciation plus imputed income. Then the following relation holds: $CE = C + CD - UD$, from which it follows that

$$\frac{\partial^2 CE}{\partial Y^2} = \frac{\partial^2 C}{\partial Y^2} + \frac{\partial^2 CD}{\partial Y^2} - \frac{\partial^2 UD}{\partial Y^2}. \tag{10}$$

It is clear from this equation that a theoretical model which implies $\partial^2 C/\partial Y^2 < 0$ carries no obvious prediction about the sign of $\partial^2 CE/\partial Y^2$; in particular, it is possible for $\partial^2 CE/\partial Y^2$ to be zero or even positive. One objective of the present research was thus to test for distribution effects using aggregate *consumption,* rather than *consumer expenditures.* Fortunately, such a time series (complete with a consistent definition of disposable income)[11] has been constructed by the builders of the MIT–Penn–SSRC (MPS) econometric model.[12]

C. *A Statistical Model Deduced from the Theory*

While the theoretical bases of Friedman's permanent income model and Modigliani-Brumberg's life-cycle model are identical, the empirical formulations differ. I have followed Friedman's model (which expresses consumption as a function of current income and lagged consumption), rather than MBA's (which expresses consumption in terms of current *labor* income and net worth) solely because of data limitations. While there are several annual time series on the distribution of income, there are no time series on the distribution of wealth, and the distribution of labor income can only be guesstimated from the available data. (Example: How do you decompose income of the self-employed into "labor" and "property" components?)

---

[10] This difficulty is noted by Mayer (1972, pp. 12–16), who nonetheless uses consumer expenditures in his tests.

[11] As Mayer (1972, p. 15) notes, this is important. The few previous studies of the consumption function which used C, rather than CE, as the dependent variable failed to use a consistent concept of disposable income.

[12] It should be noted that the MPS model does not classify residential structures as a consumer durable. Thus neither consumption nor disposable income includes the imputed yield on owner-occupied houses.

The model of Section II*A* implies that permanent consumption, c^*, is some (nonlinear and complicated) function of permanent income: $c^* = \psi(y^*)$. To make the problem tractable, I assume that $\psi(\cdot)$ is approximately linear *within* each income class. That is, if i is the index of income class, I assume $c_{it}^* = \gamma_i + k_i y_{it}^*$, where the γ_i's and k_i's may depend on interest rates. Allowing for some transitory consumption which is uncorrelated with permanent income, the expression for measured consumption in the ith income group is

$$c_{it} = \gamma_i + k_i y_{it}^* + u_{it}, \tag{11}$$

where u_{it} is the transitory component. Summing over i to obtain aggregate consumption yields:

$$c_t = \sum_i c_{it} = \gamma + \sum_i k_i y_{it}^* + v_t, \tag{12}$$

where

$$\gamma \equiv \sum_i \gamma_i \quad \text{and} \quad v_t \equiv \sum_i u_{it}.$$

Were data on permanent income by income class available, (12) could be estimated directly. However, in the absence of such data, it is necessary to have proxies for the permanent income of each group. Following Friedman's suggestion, I assume

$$y_{it}^* = (1 - \lambda_i)[y_{it} + (1 + m_i)\lambda_i y_{i,t-1} + (1 + m_i)^2 \lambda_i^2 y_{i,t-2} + \dots],$$

where m_i is an extraneously estimated growth rate and $0 \leq \lambda_i \leq 1$. Under this assumption, (11) can be expressed as:

$$c_{it} = \gamma_i[1 - (1 + m_i)\lambda_i] + k_i(1 - \lambda_i) y_{it} + (1 + m_i)\lambda_i c_{i,t-1} + \eta_{it}, \tag{13}$$

where $\eta_{it} = u_{it} - (1 + m_i)\lambda_i u_{i,t-1}$. Lacking data on consumption by income class, I am forced to assume that λ is the same for each income group. While m_i could easily be made different for each income class, the actual differences are so trivial that I ignore them. Summing (13) over i with $\lambda_i = \lambda$ and $m_i = m$ leads to

$$C_t = \gamma[1 - (1 + m)\lambda] + (1 - \lambda) \sum_i k_i y_{it} + (1 + m)\lambda C_{t-1} + \eta_t,$$

where

$$\eta_t = \sum_i \eta_{it}.$$

Note that if the model is well specified, the u_{it} in (11) will not display much serial correlation, but η_t will.[13] Also, it seems likely that the u_{it} (and hence η_t) would be heteroskedastic. If the standard deviation of the u_{it} grows proportionately with aggregate disposable income, a more efficient estimating equation would be

$$\frac{C_t}{Y_t} = \frac{\gamma^*}{Y_t} + (1 - \lambda) \sum_i k_i \left(\frac{y_{it}}{Y_t}\right) + (1 + m)\lambda \frac{C_{t-1}}{Y_t} + \varepsilon_t, \qquad (14)$$

where $\gamma^* \equiv \gamma[1 - (1 + m)\lambda]$, $\varepsilon_t \equiv \eta_t/Y_t$, and Y_t is aggregate disposable income. In words, (14) requires regressing the average propensity to consume (APC) on the inverse of disposable income, the shares of each income group,[14] and the lagged APC adjusted for growth, $C_{t-1}/Y_t = (C_{t-1}/Y_{t-1})(Y_{t-1}/Y_t)$. If each k_i also depends on the rate of interest, r_t, then interaction terms between r_t and each income share should also be included. Since with five quintile shares this would involve nearly as many coefficients as observations, the constraint that $k_i = w_i + \xi r_t$ is imposed. That is, w, but not ξ, is permitted to vary across income classes.

IV. Empirical Results

A. *The Data*

The average propensity to consume was obtained by dividing consumption (MPS definition) by disposable income (also MPS definition). Continuous time series (annually from 1947 to 1972) on the shares received by each quintile of families were obtained from the Bureau of the Census.[15]

---

[13] Worse yet, η_t is a first-order moving average rather than an autoregressive process. Strictly speaking this makes standard procedures for dealing with autocorrelation inappropriate. However, Shaman (1969) has shown that the inverse of the covariance matrix of a first-order moving average is best approximated by the inverse covariance matrix of a first-order autoregressive. This provides a pragmatic justification for my use of the Cochrane-Orcutt technique in the subsequent estimation. I am indebted to Michael Hurd for a valuable discussion on this point.

[14] Note that using the Koyck representation of permanent income makes it appropriate to use the shares in *measured*, rather than *permanent*, income for the estimation of distribution effects in (14).

The rate of interest posed the most complex measurement problem. What is wanted for a consumption function, presumably, is the *real* opportunity cost of a *consumer*. I first constructed a nominal savings rate, annually from 1949 to 1972, as a weighted average of the rates paid by commercial banks on time deposits, by savings and loan associations, and by mutual savings banks.[16] To obtain a real rate, I then subtracted a proxy for the expected rate of inflation. Following conventional procedures, I assumed an adaptive expectations mechanism:

$$\Pi_t = \theta \frac{\Delta P_t}{P_{t-1}} + (1 - \theta)\Pi_{t-1}, \tag{15}$$

where Π_t is the expected rate of inflation and P_t is the actual price level (the deflator for personal consumption expenditures). To be sure that initial conditions did not influence the series, I started the series by assuming $\Pi_t = \Delta P_t/P_{t-1}$ for $t = 1930$, used the recursion formula (15) to generate Π_t from 1931 forward, and then discarded all the data prior to 1949.[17] The value of θ was chosen from the set $[0, 0.1, 0.2, \ldots, 1.0]$ so as to minimize the standard error of each regression. This is approximately equivalent to maximizing the likelihood function over θ, and the optimal value of θ is reported (without a standard error) with each regression.

B. *Discussion of a Failure*

In brief, the regression is of the form

$$\frac{C_t}{Y_t} = \frac{a_0}{Y_t} + a_4 + a_1F_{1t} + a_2F_{2t} + a_3F_{3t} + a_5F_{5t} + a_6r_t + a_7\frac{C_{t-1}}{Y_t} + \varepsilon_t, \tag{16}$$

---

[15] I wish to thank the Bureau of the Census for furnishing me with the data prior to publication. The distributions, of course, are based on the Current Population Survey (CPS) definition of total money income. It should be noted that the shares were tabulated from ungrouped data from 1958 through 1972, but from grouped data from 1947 through 1958. In using these series I averaged the two figures for 1958. See U.S. Bureau of the Census (1973, table 16).

[16] The three constituent rates were obtained from the MPS model data file and are only available from 1949 on. The weights used were 7/16 for time deposits, 6/16 for savings and loan shares, and 3/16 for mutual savings bank shares. I adopted this weighting scheme from Springer (1973).

[17] It is hard to defend the notion that expectations are strictly adaptive in historical episodes that include drastic events such as the Great Depression and World War II. However, my estimated θ is always so high that the influence of 1930–47 price behavior on 1949–72 expectations is negligible.

where F_{it} is the share of total income received by the ith quintile (counting from the bottom) of families. Note that since

$$\sum_{i=1}^{5} F_{it} = 1$$

for all t, one quintile has to be omitted in order to avoid exact multicollinearity. The choice is arbitrary, and I selected the fourth quintile because it had the least variability.[18]

There are three obvious reasons why I was doomed to failure. First, with only 24 annual observations it is asking a lot of the data to estimate eight coefficients (seven parameters in [16] plus the autocorrelation coefficient). Second, as is well known, the income distribution has been relatively stable since World War II; and this means the F's have very modest variances. Finally, even with one share omitted, considerable collinearity remains, since $F_{1t} + F_{2t} + F_{3t} + F_{5t} = 1 - F_{4t}$, for all t, and F_{4t} is nearly constant through time.

For what it is worth, the regression is reported in column 2 of table 2. A constrained regression, which assumes equal MPCs for all income classes, is presented in column 1 for comparison. Even this latter regression allows for distribution effects of sorts. Recall that the MBF model has distribution effects unless $\delta = \beta$, and, when this equality holds, the consumption function should be strictly proportional. Since the constant (i.e., the coefficient of $1/Y$) in column 1 is significant at the 2 percent level (by a two-tail test), we can reject strict proportionality. On the basis of an extraneous estimate of the growth rate of real disposable income (FMP concept) of 3.83 percent per annum, it is possible to identify the underlying parameters. The implied estimate of λ is 0.24, a rather faster speed of adjustment than found by Friedman. Similarly, the long-run MPC (which, in this equation, is smaller than the APC) is 0.81 evaluated at the mean value of r;[19] this is, of course, lower than Friedman's estimate.

Although there is a hint of distribution effects, column 2 shows that equation (16) does not capture them at all well: the standard error of the regression exceeds that for column 1, and the estimated short-run MPCs are very suspect, ranging from 1.06 to 0.07. Clearly there is too much multicollinearity among the F_i's and too little data to get accurate estimates of subtle differences in MPCs.

One obvious approach is to omit one or more of the shares. This already compromises the theoretical model by imposing the constraint that some

---

[18] Analogous regressions were run using the shares received by unrelated individuals as well. The results were always similar, but slightly inferior to those reported below.

[19] The estimated short- and long-run MPCs are actually quite insensitive to r. The interest elasticity of consumption (evaluated at the means) in col. 1 is $-.0027$ in the short run and $-.0035$ in the long run. Other equations gave similarly trivial elasticities.

TABLE 2

REGRESSIONS WITH QUINTILE SHARES, 1949–72

VARIABLE	COEFFICIENTS (t-RATIOS) (1)	(2)	(3)*	(4)	(5)
$1/Y$	57.51 (2.6)	46.28 (2.1)	42.80 (2.1)	46.36 (2.2)	44.77 (2.3)
Constant	.616 (16.0)	.066 (0.2)	.113 (0.3)	.130 (0.4)	.264 (1.5)
r	−.0019 (−2.2)	−.0022 (−2.1)	−.0021 (−2.5)	−.0018 (−2.4)	−.0022 (−2.9)
C_{-1}/Y	.253 (3.4)	.290 (3.8)	.312 (4.1)	.312 (3.9)	.295 (4.1)
F_1	. . .	.692 (0.8)	.796 (1.1)	.837 (1.2)	.515 (1.9)
F_2	. . .	.168 (0.1)	. . .	. . .	. . .
F_3	. . .	.997 (1.2)	.908 (1.1)	.837 (1.2)	.515 (1.9)
F_5	. . .	.721 (1.5)	.645 (1.7)	.621 (1.8)	.515 (1.9)
ρ	.93 (12.1)	.92 (11.5)	.92 (11.4)	.92 (11.6)	.92 (11.5)
θ	0.8	0.8	0.9	1.0	0.9
R^2	.944	.953	.952	.952	.952
SE	.00351	.00364	.00356	.00344	.00336

NOTE.—Estimation was by the Cochrane-Orcutt iterative technique; $\hat{\rho}$ is the estimated autocorrelation coefficient; t-ratios are in parentheses; as the standard errors are only valid under the assumption that θ is known, these ratios are indicative only; θ is the weight given to actual inflation in forming inflationary expectations (see eq. [15]).

* The equation with $\theta = 0.5$ had virtually the same standard error.

quintile has the same MPC as the fourth. For want of a better criterion, I omitted the quintile whose estimated MPC was closest to that of F_4; this led to the regression reported in column 3 of table 2. The resulting standard error of estimate is still larger than in column 1 and autocorrelation is no less severe. As before, computed standard errors for individual shares are very high, suggesting multicollinearity.[20] I then imposed the further constraint that the first and third quintiles have identical MPCs to arrive at the regression in column 4. Here at last the standard error of the regression is reduced below that of the no-distribution-effects case, but the distribution variables are not very important.[21] Finally, using the same criterion, I omitted still another variable, leaving a single distributional variable in the regression. Column 5 is equation (16) subject to the constraints $a_2 = 0$, $a_1 = a_3 = a_5$.

---

[20] In view of the way the hypothesis is obtained from the data, these standard errors are purely indicative and cannot be used for hypothesis testing. In particular, the ratio of a coefficient to its "standard error" certainly does not have a t-distribution.

[21] An "F-test" of the null hypothesis that both distribution variables have zero coefficients fails to reject the null hypothesis at the 5 percent level, but just barely. The computed F-value is 3.04, as compared to a critical 5 percent point in the $F(2, 17)$ distribution of 3.59. As noted in n. 20 above, the test statistic does not really have an F-distribution because the restrictions $a_2 = 0$ and $a_1 = a_3$ were obtained from the data. The "test" is meant to be heuristic only.

Of all the regressions reported in table 2, only column 5 can really be said to be an improvement over column 1. The lone distribution variable, $F_1 + F_3 + F_5$, has a coefficient nearly twice as large as its standard error.[22] This equation implies that the short-run MPCs (evaluated at the mean interest rate) are 0.26 for the second and fourth quintiles and 0.78 for the first, third, and fifth. The corresponding long-run MPCs are 0.36 and 1.09. The pattern is not entirely believable but is probably the best job of estimating quintile-specific MPCs that can be done in the absence of quintile-specific consumption data. In any event, there is certainly no indication that MPCs decline in higher income brackets, as is commonly assumed. Precisely what they do is not illuminated very well by table 2.

C. *The First Compromise Model*

It is clear from these results that some compromise with the theory must be made if any estimation is to be done. And it is not clear that omitting variables (i.e., constraining certain MPCs to be equal) is the ideal compromise. I therefore tried two other approaches which at least have the virtue of allowing every distributional shift to affect aggregate consumption. The first involves constraining the way the MPC varies by income class and is explained in this subsection. The second entails replacing the quintile shares by one or another aggregate index of inequality and is discussed in the following subsection.

The basic model which I would like to estimate is essentially

$$c_{it} = \gamma_i + (k_0^i + k_1^i r_t) y_{it} + \lambda c_{i,t-1} + u_{it}. \qquad (17)$$

The problem is that collinearity among the y_{it} precludes accurate estimation of the k_0^i and the k_1^i. Taking a cue from the technique introduced by Almon (1965) to cope with a similar problem in the case of distributed lags, one possibility is to assume a functional form for the dependence of k_0^i and/or k_1^i on i. I report below regressions based on the assumption that both of these coefficients are linear functions of i, but I also ran equations with k's assumed to be either quadratic or logarithmic functions of i. The results were essentially identical. Appending to equation (17) the constraints $k_0^i = m_0 + m_1 i$ and $k_1^i = n_0 + n_1 i$, summing over i, and simplifying leads to

$$C_t = \gamma + m_0 Y_t + n_0 r_t Y_t + m_1 \sum_{i=1}^{5} i y_{it} + n_1 r_t \sum_{i=1}^{5} i y_{it} + \lambda C_{t-1} + V_t,$$

---

[22] The reader is again reminded of the caveat raised in n. 20 above.

where

$$\gamma = \sum_{i=1}^{5} \gamma_i, \quad V_t = \sum_{i=1}^{5} u_{it}, \quad Y_t = \sum_{i=1}^{5} y_{it}.$$

Dividing through by Y_t, and denoting the distributional variable,

$$\sum_{i=1}^{5} i\left(\frac{y_{it}}{Y_t}\right),$$

by D_t, leads to the estimating equation

$$\frac{C_t}{Y_t} = \frac{\gamma}{Y_t} + m_0 + n_0 r_t + m_1 D_t + n_1 r_t D_t + \lambda \frac{C_{t-1}}{Y_t} + v_t, \qquad (18)$$

where $v_t = V_t/Y_t$. Unfortunately, multicollinearity has still not been purged from the equation. Since D_t is relatively constant, r and rD are almost perfectly correlated, so I had to estimate one of two alternative models:

$$\frac{C_t}{Y_t} = \frac{\gamma}{Y_t} + m_0 + n_0 r_t + m_1 D_t + \lambda \frac{C_{t-1}}{Y_t} + v_t \quad (n_1 = 0); \qquad (19a)$$

$$\frac{C_t}{Y_t} = \frac{\gamma}{Y_t} + m_0 + n_1 r_t D_t + m_1 D_t + \lambda \frac{C_{t-1}}{Y_t} + v_t \quad (n_0 = 0). \qquad (19b)$$

Both variants are reported in table 3. Once again, the expectations parameter in the definition of the real interest rate (θ) was chosen to minimize the standard error; only the results with the optimal choice of θ are given in the table.[23]

It is obvious from table 3 that the choice between (19a) and (19b) is a matter of indifference. The regressions tell a story which is rather similar to that of table 2. Even if the value of θ were known a priori (so that the t-ratios reported in the table were valid), the null hypothesis of no distribution effects (i.e., the null hypothesis that the coefficient of D or rD is zero) could not be rejected at the 10 percent level (two-tail test). But the point estimate suggests that increasing inequality actually increases consumption. To give the reader some feeling for magnitudes, when both r_t and D_t are at their mean values, the predicted short- and long-run MPCs are 0.63 and 0.82, respectively. If D should then rise by 10 percent, these figures would increase to 0.67 and 0.87. While these are not dramatic changes, they are substantial relative to typical year-to-year fluctuations in the observed APC. Were the point estimate of the distributional coefficient more precise, I would be tempted to conclude that

---

[23] Other results are available on request. In view of the interest in money illusion elicited by Branson and Klevorick's paper (1969), I experimented with an alternative specification using the inverse of *nominal* disposable income in place of the inverse of *real* disposable income. In every case, the real specification gave a better fit, indicating an absence of money illusion.

TABLE 3

REGRESSIONS WITH CONSTRAINED MPCs, 1949–72

	COEFFICIENTS (t-RATIOS)	
VARIABLE	(1)	(2)
$1/Y$	49.64 (2.6)	49.51 (2.6)
Constant	.283 (1.2)	.279 (1.2)
C_{-1}/Y	.240 (3.7)	.240 (3.7)
r	−.0028 (−2.8)	...
D	.092 (1.5)	.093 (1.5)
rD	...	−.0007 (−2.8)
β	.92 (11.6)	.92 (11.6)
θ	0.7	0.7
R^2	.947	.947
SE	.00352	.00352

NOTE.—See general note to table 2.

there are moderate distributional effects which are opposite in direction to those normally assumed. However the large standard error makes this temptation easy to resist.

D. *The Second Compromise Model*

My second compromise approach is to give up on estimating separate MPCs by income class in favor of estimating the effect of income inequality on aggregate consumption directly, using some conventional measure of inequality such as the Gini ratio or the variance of the logarithms. The great advantage of this approach is, of course, that it saves on degrees of freedom without constraining any particular MPCs to be equal. The disadvantage is that it embodies a weaker "no distribution effects" assumption of its own. For example, by employing the Gini ratio as the measure of inequality, I essentially impose the constraint that all possible redistributions which raise the Gini ratio by 0.01 have the same effect on consumption. Obviously, this need not be true, but it seems more innocuous than assuming away any distribution effects whatever. In any case, I prefer to view this model as a crude approximation to the true model—an approximation dictated by the weakness of the data.

The permanent income hypothesis is generally specified for regression purposes as

$$C_t = \alpha_0 Y_t + \alpha_4 C_{t-1} + v_t. \qquad (20)$$

Friedman derived this by assuming that consumption is proportional to permanent income and that permanent income is a Koyck lag on measured

TABLE 4

REGRESSIONS WITH CURRENT INEQUALITY MEASURES

VARIABLE	INEQUALITY MEASURE σ^2 (1)	σ^2_M (2)	σ^2_F (3)	σ^2_{25} (4)	σ^2_{64} (5)	G (6)
$1/Y$	10.18	52.35	47.13	51.30	57.38	−9.48
	(1.3)	(2.2)	(2.0)	(2.1)	(2.5)	(−3.0)
Constant...	.347	.567	.594	.596	.630	.371
	(7.4)	(11.9)	(9.4)	(12.8)	(17.3)	(4.1)
r..........	−.0030	−.0022	−.0022	−.0021	−.0020	−.0035
	(−3.5)	(−2.6)	(−2.6)	(−2.6)	(−2.51)	(−3.5)
C_{-1}/Y	.509	.290	.307	.293	.244	.631
	(7.0)	(3.4)	(3.7)	(3.6)	(3.4)	(10.9)
d	.083	.012	−.014	−.004	.012	.072
	(2.5)	(0.5)	(−0.4)	(−0.1)	(2.1)	(0.3)
β	.19	.92	.92	.93	.94	.14
	(0.9)	(11.0)	(11.0)	(11.4)	(12.8)	(0.6)
θ	1.0	0.9	0.9	0.9	0.8	1.0
R^2	.942	.946	.946	.956	.958	.947
SE........	.00381	.00366	.00367	.00338	.00313	.00392

NOTE.—*See general note to table 2. The periods of estimation are: for* σ^2, σ^2_M, *and* σ^2_F, *1949–70; for* σ^2_{25} *and* σ^2_{64}, *1949–69; for* G, *1949–52, 1954–68.*

income. But since this is the general framework in which I have tested for distribution effects throughout, it is worth noting that (20) can arise in other models as well. For example, (20) could represent Brown's (1952) habit-persistence model. Also, a regression very close to (20) could represent Duesenberry's (1949) relative income hypothesis, since, in annual data for the postwar era, "previous peak income" and lagged income are almost always identical.

The pure theory of consumer behavior implies that α_0 should be a function of the rate of interest, r_t, and I simply propose to add the inequality in the size distribution of income, d_t, to the list of arguments. That is, $\alpha_0 = \alpha_1 + \alpha_2 d_t + \alpha_3 r_t$. The null hypothesis to be tested is $\alpha_2 = 0$ against the two-tailed alternative: $\alpha_2 \neq 0$. The regression to be estimated, then, is[24]

$$C_t = \gamma + (\alpha_1 + \alpha_2 d_t + \alpha_3 r_t) Y_t + \alpha_4 C_{t-1} + v_t. \tag{21}$$

---

[24] Equation (21) cannot be offered as an accurate representation of Friedman's model for the following reason. Adding distribution effects to Friedman's model in the way I have suggested gives $C_t = (k_0 + k_1 d_t + k_2 r_t) Y_t^*$, where Y^* is permanent income. Adopting the Koyck lag for permanent income, as Friedman suggested, gives $Y_t^* - \lambda Y_{t-1}^* = (1 - \lambda) Y_t$, but applying the same Koyck transformation to C_t gives

$$\begin{aligned} C_t - \lambda C_{t-1} &= k_t Y_t^* - \lambda k_{t-1} Y_{t-1}^* \\ &= k_t (Y_t^* - \lambda Y_{t-1}^*) + \lambda Y_{t-1}^* (k_t - k_{t-1}) \\ &= k_{t(1-\lambda)} Y_t + \lambda Y_{t-1}^* (k_1 d_t - k_1 d_{t-1} + k_2 r_t - k_2 r_{t-1}). \end{aligned}$$

Since Y_{t-1}^* is not observable, this equation is not suitable for empirical analysis. Thus eq. (21) was adopted instead. Note that this difficulty would arise even without distribution effects, as long as the rate of interest is allowed to affect the MPC.

TABLE 5

REGRESSIONS WITH LAGGED INEQUALITY MEASURES

	INEQUALITY MEASURE					
VARIABLE	σ^2 (1)	σ^2_M (2)	σ^2_F (3)	σ^2_{25} (4)	σ^2_{64} (5)	G (6)
$1/Y$	16.98	35.27	47.28	−8.94	−9.70	53.24
	(2.0)	(2.2)	(2.2)	(−3.1)	(−4.2)	(2.3)
Constant	.353	.521	.562	.395	.393	.641
	(8.5)	(10.7)	(8.6)	(7.3)	(8.5)	(7.8)
r	−.0021	−.0026	−.0022	−.0032	−.0034	−.0070
	(−2.2)	(−3.2)	(−2.7)	(−3.4)	(−3.8)	(−3.2)
C_{-1}/Y	.457	.340	.311	.580	.588	.220
	(5.6)	(5.5)	(4.0)	(8.0)	(9.3)	(3.4)
d_{-1}	.099	.034	.010	.060	.071	.066
	(3.1)	(1.5)	(0.3)	(1.8)	(2.8)	(0.4)
$\hat{\rho}$	.07	.91	.92	.29	.19	.93
	(0.4)	(10.2)	(11.4)	(1.4)	(0.8)	(10.6)
θ	1.0	0.8	0.9	1.0	1.0	0.3
R^2	.940	.948	.947	.928	.940	.961
SE	.00382	.00353	.00359	.00414	.00376	.00334

NOTE.—See general note to table 2. The periods of estimation are: for σ^2, σ^2_M, and σ^2_F, 1949–71; for σ^2_{25} and σ^2_{64}, 1950–70; for G, 1949–53, 1955–69.

I again correct for heteroskedasticity by dividing (21) through by Y_t and estimate

$$\frac{C_t}{Y_t} = \frac{\gamma}{Y_t} + \alpha_1 + \alpha_2 d_t + \alpha_3 r_t + \alpha_4 \frac{C_{t-1}}{Y_t} + e_t, \tag{22}$$

where $e_t = v_t/Y_t$.

Several time series on overall income inequality, d_t, are available. Since it is by no means clear which measure best captures the relevant distributional shifts, I have run regressions with each of them. The measures are:

G = the Gini concentration ratio of the distribution of money income (CPS concept) among families and unrelated individuals. This series is available for 1948–68 (with 1953 missing) in Budd (1970).

σ^2 = the variance of the logarithms of CPS money income among all persons with income over 14 years of age. This is available over 1947–70, as computed by Schultz (1971).

σ^2_M = same as σ^2, but restricted to males.

σ^2_F = same as σ^2, but restricted to females.

σ^2_{25} = same as σ^2_M, but restricted to men at least 25 years of age. This has been calculated over 1949–69 by Chiswick and Mincer (1972).

σ^2_{64} = same as σ^2_{25}, but excluding men over 65.

Following a suggestion made by Lubell (1947), I tried each variant of d in both current and lagged form. Regressions using current d are reported in table 4, and regressions using lagged d are reported in table 5.

Note that the period of estimation differs somewhat, depending on which variant of d is employed. Because of the paucity of data, I used every available data point rather than confine myself to a common sample period (which would have been 1949–52, 1954–68).

If σ^2, the variance of logarithms over the entire adult population, is used as the inequality measure, it does not matter much whether d_t or d_{t-1} enters the regression. In either the current or the lagged version the impact of inequality is apparently measured with some precision, and an increase of 0.03 in σ^2 (which is a fairly typical year-to-year change) would increase the average propensity to consume by about 0.3 of a percentage point in the short run and about 0.5 of a percentage point in the long run. These two equations are also notable for the absence of autocorrelation (a rare finding in this study) and for the slow speed of adjustment. In fact, inspection of the tables reveals a systematic relationship: the equations with slow adjustment speeds do not have autocorrelated residuals.

The only other measure which is "significant"[25] in both the current and the lagged specification is σ^2_{64}, the log variance among males aged 25–64. The current version exhibits trivially small distribution effects, and the lagged version has much larger ones. Other than these, only the regression with lagged σ^2_{25} indicates significant distribution effects (at the 10 percent level in a two-tailed test). However, the persistent sign pattern is suggestive. Except for two trivially small coefficients, the point estimates all say that a rise in an inequality index leads to higher consumption. The next section is devoted to interpreting this conclusion and convincing the reader that it is not quite so outlandish as it may seem.

V. Can the Results Be Right?

I began this paper by contrasting what might be called the educated layman's view (that more equal income distributions give rise to more consumption) with the view that is now dominant among macroeconomists (that the income distribution does not matter). The empirical results certainly contradict the layman's view. Instead, they suggest either that consumption is independent of the income distribution or that distributions with less measured inequality give rise to somewhat less consumption. Is the latter possibility believable?[26]

To begin with pure theory, I showed in Section II*A* that—in the optimal life-cycle consumption model—transfers from poor to rich will actually increase consumption if the elasticity of the marginal utility of bequests, β, exceeds the elasticity of the marginal utility of consumption,

---

[25] As noted, significance tests are valid only on the assumption that θ is known a priori.

[26] To be sure, the arguments I am about to give are not terribly convincing on a priori grounds. They are offered as conceivable explanations of a counter-intuitive result which gets at least mild support from the data.

TABLE 6

POSTWAR CHANGES IN INCOME INEQUALITY

Measure	Initial Value*	Final Value†
σ^2	0.785	1.406
σ^2_M	0.668	1.187
σ^2_F	0.670	1.169
σ^2_{25}	0.742	0.729
σ^2_{64}	0.653	0.581
G	0.424	0.406

* 1947 for σ^2, σ^2_M, and σ^2_F; 1948 for G; and 1949 for σ^2_{25} and σ^2_{64}.
† 1968 for G; 1969 for σ^2_{25} and σ^2_{64}; 1970 for σ^2, σ^2_M, and σ^2_F.

δ. So one interpretation of the data—an interpretation which I do not find particularly appealing—is that the rich actually consume a larger fraction of their lifetime resources because bequests have a wealth elasticity less than unity.[27]

Duesenberry's relative income hypothesis gives an alternative theoretical rationale for the empirical findings. In his model, utility attaches not to consumption but to the *ratio* of own consumption to a weighted average of consumption of others. The weights reflect the frequency of contact with individuals in other consumption classes, and Duesenberry hypothesizes that more contacts with individuals with higher consumption will increase the fraction of income that is consumed. Thus, *it is possible* that an equalization of the income distribution could reduce the number of contacts which most people have with persons much better off than themselves, and therefore reduce aggregate consumption (Duesenberry 1949, pp. 44–45).[28]

A third explanation of the findings, and the one I find most satisfying, rests on the distinction between the kind of "ideal" redistributions that pure theory envisions and the actual redistributions that are reflected in postwar United States data. Table 6 shows the net change over the entire sample period in each of the six measures of income inequality. The variable G is conceptually different from the other variables in two ways. First, it is a Gini ratio, not a log variance. But, more important, it uses families and unrelated individuals (pooled) as the recipient unit, whereas all the others are based on individuals. Thus, the six measures tell the following story. The distributions of income among *families*, and among males over 24 years old, hardly changed in the postwar period; the decline in inequality was very slight. However, inequality fell much more

---

[27] Becker's (1974) analysis shows that $\delta \leq \beta$ implies that the intergenerational distribution of consumption shifts in favor of the current generation as the economy gets richer. See n. 3 above.

[28] While this is possible, the reverse could also happen, as was pointed out by Johnson (1951).

TABLE 7

FRACTIONS OF ALL PERSONS RECEIVING INCOME, BY AGE-SEX GROUP

Age Group	Males 1948	Males 1972	Females 1948	Females 1972
14–19	3.8	5.7	3.0	4.9
20–24	7.2	6.6	4.3	5.8
25–34	15.4	11.0	6.6	7.4
35–44	14.0	8.9	6.1	6.1
45–54	11.7	9.1	4.8	6.5
55–64	8.9	7.2	3.5	5.8
65 and up	6.6	6.8	4.1	8.3

SOURCE.—Computed by the author from data in U.S. Bureau of the Census (1967, table 14; 1973, table 47).
NOTE.—Totals may not add to 100% as a result of rounding.

noticeably among prime-age males (25–64 years old)[29]—suggesting that the gap between old and prime-age men widened. Further, among all individuals above 14 years of age, inequality rose substantially in the total population, among males, and among females. This suggests that substantial increases in the labor force participation of young people and women may have raised inequality by adding many new income recipients to the lower tail of the distribution.

A more detailed look at these phenomena can be obtained by consulting tables 7–10. Tables 7 and 8 analyze the age-sex composition of the population which underlies each of the income distribution measures other than G, that is, all persons with income. It is clear from table 7 that the distribution included many more teenage boys, and females of all ages except 35–44, in 1972 than it did in 1948. There are two principal reasons for such changes over time: demographic shifts in the age-sex composition of the population as a whole, and changes in labor force participation rates.[30] Table 8 shows that it is the former that accounts for the increased importance of teenage boys; their participation rate actually fell. Similarly, exogenous demographic factors appear to play the major role in the increased numbers of very young and very old women having income. However, for women aged 20–64, it appears that their increased importance in the income distribution is largely attributable to higher labor force participation.

Tables 9 and 10 contain similar information for shares in *income*, rather than in population. Table 9 highlights the increased importance of women over 20 in the overall income distribution. While they received only 17.2 percent of all income in 1948, this share rose to 24.4 percent by

---

[29] Shultz (1971, p. 11) actually finds a 17 percent *rise* in the log variance of income among males aged 25–64 over the 1947–70 period. But most of this occurs between 1947 and 1949.

[30] This ignores any changes in the fraction of each age-sex group receiving property (but not labor) income.

TABLE 8

LABOR FORCE PARTICIPATION RATES, BY AGE-SEX GROUP

	MALES		FEMALES	
AGE GROUP	1948	1972	1948	1972
14–19	54.2	46.9	32.7	35.8
20–24	85.7	85.9	45.3	59.1
25–34	96.1	95.9	33.2	47.6
35–44	98.0	96.5	36.9	52.0
45–54	95.8	93.3	35.0	53.9
55–64	89.5	80.5	24.3	42.1
65 and up	46.8	24.4	9.1	9.3

SOURCE.—Figures for 14–19 age group were computed by the author from data in U.S. Office of the President (1974, table A-1). Other figures came directly from that source.

TABLE 9

SHARES IN TOTAL MONEY INCOME, BY AGE AND SEX

	MALES		FEMALES	
AGE GROUP	1948	1972	1948	1972
14–19	1.1	1.3	0.9	0.8
20–24	6.0	5.2	2.5	3.0
25–34	19.4	16.7	4.2	5.0
35–44	21.4	17.0	4.0	4.3
45–54	17.2	17.4	3.3	4.8
55–64	11.5	11.7	1.8	3.7
65 and up	5.1	5.7	1.4	3.6

SOURCE.—Computed by the author from data in U.S. Bureau of the Census (1967, tables 14, 36; 1973, table 47).
NOTE.—Totals do not add to 100% as a result of rounding.

TABLE 10

RELATIVE MEAN INCOMES, BY AGE AND SEX*

	MALES		FEMALES	
AGE GROUP	1948	1972	1948	1972
14–19	0.20	0.12	0.20	0.09
20–24	0.54	0.42	0.38	0.27
25–34	0.83	0.79	0.41	0.35
35–44	1.00	1.00	0.42	0.37
45–54	0.95	0.99	0.44	0.39
55–64	0.84	0.85	0.35	0.23
65 and up	0.51	0.44	0.23	0.29

SOURCE.—Same as for table 7.
* Ratio of mean income in each group to mean income of males aged 35–44.

1972. And table 10 shows that this was accomplished despite a widening in the relative income gap between men and women. In fact, it is apparent from table 10 that all groups except elderly women lost ground relative to prime-age males during this period.

In a word, these tables suggest the following anatomy of the rise in *measured* inequality *among persons with income* over the postwar period:

1. Purely demographic forces[31] led to a substantial increase in the number of teenagers and elderly women receiving income. Since these groups generally have low mean incomes (and, in the case of teenagers, also relatively high within-group variance), this alone would tend to raise inequality. Compounding this, the mean income of teenagers relative to prime-age males fell so sharply that their share in total income actually declined.[32]
2. Marked increases in labor force participation of women between 20 and 64 years of age added a great number of relatively low incomes to the distribution, thus raising inequality as conventionally measured. This alleged rise in inequality strikes me as particularly illusory, since it is attributable to an artifact in construction of the data, that is, to defining the population as "all persons with income" rather than "all persons." Presumably, if all the "zeros" in both years had been included in the income distribution, the growth of the female labor force would have *reduced* measured inequality.
3. Incomes of males aged 20–24 and over 64, as well as all but the oldest women, declined relative to prime-age males. This represents a bona fide increase in inequality by most reasonable criteria, but it is very far afield from the idealized "regressive transfers" discussed in Section II*B*.

If these were the underlying changes in the income distribution, the regressions in Section IV*D* could be associating greater inequality with greater consumption if any of the following are true: (*a*) elderly women have higher-than-average marginal propensities to consume, (*b*) women have greater MPCs than men, or (*c*) prime-age males have lower-than-average MPCs.

Of these possibilities, the theory of optimal life-cycle consumption gives every reason to believe that (*a*) and (*c*) would be true. Given a typical "humped" income profile, consumer units will dissave while very young and very old, and save during the prime earning years. However, there seems to be no theoretical reason to believe in the veracity of (*b*).[33] Indeed, the theory of the household gives every reason to believe that the consumption behavior of married women and married men should be essentially identical.[34] And it is the married women (in particular, those married with spouse present) who have registered the greatest gains in labor force participation over this period.[35]

---

[31] This phrase is not meant to deny that there may have been economic reasons behind these phenomena. It simply connotes that the variable itself (age-sex distribution) is a demographic one rather than an economic one (like labor-force participation rates).

[32] The drop in relative incomes may well have been caused by the increase in relative labor supplies, but that is beyond the scope of this paper.

[33] This, of course, does not prove that (*b*) is false. It would be true, for example, if wives typically entered the labor force to finance the acquisition of specific consumer durables which the family wished to purchase.

Finally, the only other study known to me which included the *size* distribution of income in the consumption function also obtained the "odd" result that increased inequality led to increased consumption. Metcalf (1972) was rather puzzled by the finding and noted that "while a number of significant relationships were uncovered, it is not yet clear how the results should be interpreted" (p. 148–49). The distributional variable in his preferred consumption function is the ratio of income at the ninetieth percentile to mean income, and he concluded that "the higher the top decile income relative to the mean, the higher the marginal propensity to consume" (p. 152).[36]

VI. Summary

In this paper I have shown that the established theory of consumer behavior carries definite implications as to the effect of a sequence of regressive transfers on aggregate consumption. Such an increase in inequality must reduce *consumption* if bequests are a luxury good (or increase consumption if own consumption is the luxury good). However, this does *not* say that such a redistribution would necessarily reduce *consumer expenditures*. Nor does it say that aggregate consumption must fall as a result of the kinds of redistributions that have taken place in the postwar United States. Finally, the theory (and the facts) give no reason to believe that a shift in the *factor share* distribution will have any particular effect upon consumption.

The only rigorously correct way to test for the existence of distribution effects in the aggregate consumption function is to estimate directly separate marginal propensities to consume by income class. Unfortunately, the data are too weak to allow this, so several "second best" procedures were adopted. First, various MPCs were constrained to be equal to one another; then a method similar to the Almon lag technique was used to constrain the variation in MPCs; finally, the MPCs were ignored and an aggregate measure of inequality was inserted in the consumption function. The upshot of all this appears to be that equalizing the income distribution will either have no bearing on or (slightly) reduce aggregate consumption.

---

[34] In March 1972, 58.4 percent of the females in the labor force were married with a husband present. The remainder included single, widowed, divorced, and separated women (see U.S. Office of the President 1974, table B-1).

[35] Labor force participation among "married, spouse present" women rose from 22 percent in 1948 to 41.5 percent in 1972, while participation rates were virtually trendless for other categories of women. (See U.S. Office of the President 1974, table B-2.)

[36] A further possibility, which I regard more as an intellectual curiosum than as a practical explanation of the results, is that the log variance could conceivably *increase* while inequality is actually falling by the mean-preserving spread criterion. To show this, I follow Atkinson's (1970) approach to inequality measurement, which assumes an additive social welfare function, $W = \int u(y) f(y)\, dy$, where $u(y)$ is the social welfare significance of a person's receiving income y, and $f(y)$ is the income density function. The specific utility function implicit in using the variance of logarithms, $V = \int (\log y - k)^2 f(y)\, dy$, where $k \equiv E(\log y)$, as the inequality measure is clearly $u(y) = (\log y - k)^2$. As first noted by Atkinson (1970, p. 13), this function is not concave over its entire range; therefore a sequence of *regressive* transfers might actually *raise* social welfare (Rothschild and Stiglitz 1973). Since five of the six inequality measures are variances of logarithms, they may not be correct indicators of the direction of change in inequality.

Several reasons for the latter possibility were suggested. Of these, I find two most appealing. First, if the kinds of "demonstration effects" stressed by Duesenberry are at all important, disequalization can conceivably lead to more rather than less consumption. Second, income inequality in the postwar United States increased largely because of demographic shifts and increased labor force participation of women. Had these women been counted as "zeros" in the income distribution while out of the labor force, rather than omitted, income inequality might well have fallen. Thus the observed positive effect of certain inequality measures on consumption may be misleading.

As a by-product, this study has shed some additional light on other properties of the consumption function. In all specifications, I find an absence of money illusion, a very small negative interest elasticity, and a rather fast adjustment of inflationary expectations to actual inflation.

Appendix

Solution of the Optimal Consumption Problem with a Bequest Motive

The problem is to pick a time pattern of consumption, $c(t)$, and a level of terminal assets, $K(T)$, so as to maximize

$$\int_0^T U[c(t)]e^{-\rho t}\,dt + B[K(T)], \tag{A1}$$

subject to the lifetime budget constraint

$$\int_0^T c(t)e^{-rt}\,dt + K(T)e^{-rT} = W. \tag{A2}$$

Defining the functional,

$$\begin{aligned} L[c(t), K(T)] = {} & \int_0^T U[c(t)]e^{-\rho t}\,dt + B[K(T)] \\ & + \lambda\left[W - \int_0^T c(t)e^{-rt}\,dt - K(T)e^{-rT}\right], \end{aligned}$$

the first-order conditions are[37]

$$\frac{\partial L}{\partial c(t)} = U'[c(t)]e^{-\rho t} - \lambda e^{-rt} = 0 \text{ for all } t; \qquad \text{(A3)}$$

$$\frac{\partial L}{\partial K(T)} = B'[K(T)] - \lambda e^{-rt} = 0. \qquad \text{(A4)}$$

Equation (A4), of course, is simply the transversality condition, since $\lambda e^{-rT} = U'[c(T)]e^{-\rho T}$ by (A3). Solving (A3) under the specific functional form $U[c(t)] = [c(t)^{1-\delta}]/(1 - \delta)$ gives

$$c(t)^{-\delta} = \lambda e^{-(r-\rho)t}, \qquad \text{(A5)}$$

from which it follows that λ is related to the initial level of consumption by

$$c(0)^{-\delta} = \lambda, \qquad \text{or } c(0) = \lambda^{-1/\delta}. \qquad \text{(A6)}$$

Therefore (A5) becomes

$$c(t) = \lambda^{-1/\delta}e^{[(r-\rho)/\delta]t} = c(0)e^{[r-\rho)/\delta]t}$$

by (A6), which is equation (3) in the text.

Now use (A6) and the specific functional form

$$B[K(T)] = \frac{bK(T)^{1-\beta}}{1 - \beta}$$

to express (A4) as $bK_T^{-\beta} = c(0)^{-\delta}e^{-rT}$, or $K_T = (be^{rT})^{1/\beta}c(0)^{\delta/\beta}$, which is equation (5) in the text.

Finally, return to the budget constraint, equation (A2), to write

$$\begin{aligned} W - K(T)e^{-rT} &= \int_0^T c(t)e^{-rt}\,dt \\ &= c(0)\int_0^T e^{[(r-\rho)/\delta]t}e^{-rt}\,dt \\ &= c(0)\int_0^T e^{([r(1-\delta)-\rho]/\delta)t}\,dt, \end{aligned}$$

which is equation (4) in the text with

$$\phi(r, \rho, \delta, T) \equiv \left[\int_0^T e^{([r(1-\delta)-\rho]/\delta)t}\,dt\right]^{-1}.$$

References

Almon, S. "The Distributed Lag between Capital Accumulations and Expenditures." *Econometrica* 33 (January 1965): 178–96.

Atkinson, A. B. "On the Measurement of Inequality." *J. Econ. Theory* 2 (September 1970): 244–63.

---

[37] Assuming $U(\cdot)$ and $B(\cdot)$ are strictly concave, these are also sufficient. Also the assumptions

$$\lim_{c(t)\to 0} U'[c(t)] = \infty \qquad \text{and} \qquad \lim_{K(T)\to 0} B'[K(t)] = \infty$$

rule out the possibility of corner solutions.

Becker, G. S. "A Theory of Social Interactions." *J.P.E.* 82, no. 6 (November/December 1974): 1063–93.

Blinder, A. S. *Toward an Economic Theory of Income Distribution.* Cambridge, Mass.: M.I.T. Press, 1974.

Branson, W. H., and Klevorick, A. K. "Money Illusion and the Aggregate Consumption Function." *A.E.R.* 59 (December 1969): 832–49.

Brown, T. M. "Habit Persistence and Lags in Consumer Behavior." *Econometrica* 20 (July 1952): 355–71.

Budd, E. C. "Postwar Changes in the Size Distribution of Income in the U.S." *A.E.R.* 60 (May 1970): 247–60.

Chiswick, B. R., and Mincer, J. "Time Series Changes in Personal Income Inequality in the United States from 1939, with Projections to 1985." *J.P.E.* 80, no. 3, pt. 2 (May/June 1972): S34–66.

Duesenberry, J. S. *Income, Saving, and the Theory of Consumer Behavior.* Cambridge, Mass.: Harvard Univ. Press, 1949.

Friedman, M. *A Theory of the Consumption Function.* Princeton, N.J.: Princeton Univ. Press (for Nat. Bur. Econ. Res.), 1957.

Goldsmith, R. *A Study of Saving in the United States.* Vol. 1. Princeton, N.J.: Princeton Univ. Press (for Nat. Bur. Econ. Res.), 1955.

Johnson, H. G. "A Note on the Effect of Income Redistribution on Aggregate Consumption with Interdependent Consumer Preferences." *Economica* 18 (August 1951): 295–97.

Kuznets, S. *Uses of National Income in Peace and War.* New York: Nat. Bur. Econ. Res., 1942.

Lubell, H. "Effects of Redistribution of Income on Consumers' Expenditures." *A.E.R.* 37 (March 1947): 157–70.

Mayer, T. *Permanent Income, Wealth, and Consumption.* Berkeley: Univ. California Press, 1972.

Metcalf, C. E. *An Econometric Model of the Income Distribution.* Chicago: Markham, 1972.

Modigliani, F., and Ando, A. "The 'Permanent Income' and the 'Life Cycle' Hypothesis of Saving Behavior: Comparisons and Tests." In *Proceedings of the Conference on Consumption and Saving,* edited by I. Friend and R. Jones. Vol. 2. Philadelphia: Univ. Pennsylvania, 1960.

Modigliani, F., and Brumberg, R. "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data." In *Post-Keynesian Economics,* edited by K. K. Kurihara. New Brunswick, N.J.: Rutgers Univ. Press, 1954.

Projector, D. S.; Weiss, G. S.; and Thoresen, E. T. "Composition of Income as Shown by the Survey of Financial Characteristics of Consumers." In *Six Papers on the Size Distribution of Wealth and Income,* edited by L. Soltow. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1969.

Rothschild, M., and Stiglitz, J. E. "Increasing Risk. I. A Definition." *J. Econ Theory* 2 (September 1970): 225–43.

———. "Some Further Results on the Measurement of Inequality." *J. Econ. Theory* 6 (April 1973): 188–204.

Schultz, T. P. "Long Term Changes in Personal Income Distribution: Theoretical Approaches, Evidence and Explanations." Paper presented at the meetings of the American Economic Association, New Orleans, December 1971.

Shaman, P. "On the Inverse of the Covariance Matrix of a First Order Moving Average." *Biometrika* 56 (1969): 595–600.

Springer, W. L. "Windfalls, Temporary Income Taxes, and Consumption Behavior." Ph.D. dissertation, Princeton Univ., 1973.

Strotz, R. H. "Myopia and Inconsistency in Dynamic Utility Maximization." *Rev. Econ. Studies* 23, no. 3 (1955–56): 165–80.

Suits, D. B. "The Determinants of Consumer Expenditure: A Review of Present Knowledge." In *Impacts of Monetary Policy*, Commission on Money and Credit. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

Taylor, L. D. "Saving out of Different Types of Income." *Brookings Papers Econ. Activity* 2, no. 2 (1971): 383–407.

U.S. Bureau of the Census. *Trends in the Income of Families and Persons in the United States, 1947-1964*. Technical Paper no. 17. Washington: Government Printing Office, 1967.

———. *Current Population Reports*. Ser. P-60, no. 90. *Money Income in 1972 of Families and Persons in the United States*. Washington: Government Printing Office, 1973.

U.S. Office of the President. *Manpower Report of the President*. Washington: Government Printing Office, 1974.

Yaari, M. E. "On the Consumer's Lifetime Allocation Process." *Internat. Econ. Rev.* 5 (September 1964): 304–17.

The Citation Practices of Doctorates in Economics

George J. Stigler and Claire Friedland

University of Chicago

The citations of doctorates in economics (who received the degree between 1950 and 1955) from six major universities are analyzed for articles published in the period 1950–68 in the two fields of value theory (microeconomics) and monetary and fiscal theory (macroeconomics). The citations are examined for commonality in the use of authorities and for parochialism with regard to the faculty of their Ph.D.-granting institution. The citation practices of the "authorities" cited by the doctorates are also examined, compared with those of the doctorates, and used to define scientific associates and adversaries.

Schools of thought have been a popular concept—or perhaps one should say, label—in economic discussion. The Physiocrats, the Ricardians, the Austrians, the Marshallians, and the Keynesians are famous examples, but we are almost as willing to speak of smaller groups such as the schools of workable competition, public choice, and Chicago.

Yet the degree of harmony in the views of a group of men necessary for them to qualify as a school is not easily specified. As a field develops, a large area of agreement—made famous by Kuhn as the paradigm—comes to be shared by all qualified practitioners, whatever their "school."

Conversely, a school of scientific thought cannot be expected to display the specificity of belief and the intolerance of dissent that a religious body or a policy-oriented group will insist upon. After all, a science pays homage to originality and to independent verification, and these are engines of change. Still, in closely knit schools it is possible to make excellent predictions of a person's views on certain controversial issues simply by knowing that he is a member.

This paper began with the question: Are the major centers of graduate instruction in the United States "schools" in the sense of leaving distinctive

[*Journal of Political Economy*, 1975, vol. 83, no. 3]
© 1975 by The University of Chicago. All rights reserved.

TABLE 1

INSTITUTIONS, DOCTORATES (1950–55), AND ARTICLES (1950–68) COMPRISING THE UNIVERSE FOR THIS STUDY

INSTITUTION	TOTAL DOCTORATES	DOCTORATES PUBLISHING ARTICLES IN		ARTICLES PUBLISHED IN	
		Value Theory	Monetary Theory	Value Theory	Monetary Theory
Berkeley	80	16	14	32	23
Chicago	106	26	27	82	73
Columbia	116	19	15	61	35
Harvard	225	38	32	127	77
M.I.T.	30	8	*	17	*
Wisconsin	143	*	8	*	19

* Too few for analysis.

imprints upon their doctorates? More precisely, does the graduate of Chicago differ in significant scientific respects from the graduate of Harvard? Even when we answer this second question, the source of any observed differences—whether instilled by graduate schools or due to basic attitudes which influenced their choice of graduate schools—will be largely outside our purview.

The basic approach was as follows. Lists of articles in two main areas of economic theory—value theory (as we shall term the theory of prices and resource allocation) and monetary and fiscal economics (often abbreviated by us as monetary theory)—written by doctorates from each of six major schools were compiled for the period, 1950–68[1] (table 1). Even the fifth largest school we included in each area has so few doctorates writing in these two fields that we are forced to skate on statistically thin ice.

In the course of our study other aspects of the citation practices of economists naturally arose, and even though our sample was not designed to deal with them, some preliminary inquiries were made. The most important of these additional topics is the citation practices of leading economists.

We begin with a survey of the academic backgrounds of these doctorates and their subsequent professional careers. After an examination of some problems in the use of citations, we shall then turn to our central question: the influence of choice of school on the doctorates. Thereafter we examine the relationships among the cited leading economists revealed by *their* citations.

---

[1] One Chicago influence is already evident. Many economists would refer to these as micro- and macroeconomic theory, respectively.

TABLE 2

TYPES OF UNDERGRADUATE SCHOOLS ATTENDED BY THE DOCTORATES
(PERCENTAGE DISTRIBUTION)

Type of Undergraduate Institution	Berkeley	Chicago	Columbia	Harvard	M.I.T.	Wisconsin	Total
United States:							
Private	21.7	53.8	44.2	55.1	43.3	19.6	41.6
Public	68.1	30.2	40.7	30.9	26.7	63.8	43.0
Total U.S.	89.8	84.0	84.9	86.0	70.0	83.4	84.6
Canada	5.8	7.5	2.7	6.3	13.3	3.6	5.6
Other foreign:							
Europe	4.3	3.8	7.1	1.0	6.7	1.4	3.2
Asia	0.0	2.8	2.6	3.4	6.7	5.8	3.5
Total foreign	4.3	8.5	12.4	7.7	16.7	13.0	9.8
Total known	100.0	100.0	100.1	100.0	100.0	99.9	100.0
Unknown*	13.8	0.0	2.6	8.0	0.0	3.5	5.3

* Percent of all doctorates.

I. The Doctorates

Our necessarily brief survey of the characteristics of the doctorates included in our study will often consider all 1950–55 doctorates from the chosen schools, not only those writing in the two subject areas included in the citation study.[2]

Who Are the Doctorates?

We begin with the baccalaureate origins of the doctorates. In general the graduate public institutions recruit primarily from public institutions and the graduate private institutions primarily from private institutions (see table 2). The segregation goes farther. In every case except M.I.T. the leading source of a graduate school is its own undergraduate college (see table 3). Perhaps the most striking characteristic of the recruitment patterns is that geographical proximity could explain a large fraction of the pairings of undergraduate and graduate schools.

The graduate schools differ substantially in the distribution of students among major fields, so the restriction of our study to articles in value

---

[2] The biographical record on the doctorates—where they came from and where they are now—is seriously incomplete. In particular, the doctorates who never publish (a third of the total, as we shall see) and those who are not members of the American Economic Association (AEA) are poorly reported.

TABLE 3

LEADING INDIVIDUAL UNDERGRADUATE SCHOOLS ATTENDED BY DOCTORATES

Undergraduate School	Berkeley	Chicago	Columbia	Harvard	M.I.T.	Wisconsin	Total
Harvard	1	5	3	39*	1	0	49*
Wisconsin	2	2	2	5	0	29	40
Chicago	1	25	2	2	0	0	30
Berkeley	15	2	1	2	1	0	21
CCNY	0	1	17	2	0	1	21
Columbia	1	1	17	1	0	0	20
Univ. of Washington	3	2	2	4	0	2	13
Univ. of Illinois....	0	0	3	3	0	3	9
Iowa State	2	0	1	2	1	3	9
Univ. of Minnesota.	2	2	4	0	0	1	9
Univ. of Texas.....	0	0	5	3	0	1	9
Oberlin	1	1	1	5	0	0	8
NYU	0	1	3	2	0	1	7
Northwestern	0	4	1	1	1	0	7
Princeton	0	3	1	2	1	0	7
Brooklyn	0	1	3	1	0	1	6
Ohio State	1	3	0	1	1	0	6
Cornell	0	2	1	2	1	0	6
Swarthmore	1	0	0	4	1	0	6
Univ. of Arkansas..	0	0	0	2	0	3	5
Kansas State	0	1	0	2	0	2	5
Queens College	0	1	3	1	0	0	5
Amherst	0	1	1	3	0	0	5
Rochester	1	0	0	2	1	1	5
Pomona	2	0	1	1	1	0	5

* Includes five Radcliffe B.A.'s.

theory and monetary and fiscal theory has a large influence upon the share of doctorates of each school represented in the citation study. The distribution of doctorates by the primary fields in which they work is shown in table 4.[3]

In 1969 the differences among schools in the type of employment of doctorates was surprisingly large (table 5). Only one Berkeley doctorate in 30 entered government service, but almost one in five from Harvard did so. A fourth of M.I.T.'s small number of doctorates entered business, against only one-fourteenth for Berkeley, Chicago, and Wisconsin.

The differences among schools would have been modified if our information had been more complete. Our chief source has been the *Handbook* of the American Economic Association (AEA), but this has been supplemented by alumni records of several universities. This latter

---

[3] The classification is based upon self-description in the 1969 *Handbook* of the AEA, supplemented by assignments on the basis of publications and 1964 *Handbook* self-descriptions. The categories of "Value Theory" and "Monetary and Fiscal Theory" are described in the *Handbook*. Figures are percentages of all doctorates with information available.

TABLE 4

DOCTORATES BY PRIMARY FIELD
(PERCENTAGE OF SCHOOL TOTALS)

	PRIMARY FIELD	
SCHOOL	Value Theory	Monetary and Fiscal Theory
Berkeley	23.4	3.1
Chicago	20.2	19.1
Columbia	20.0	14.4
Harvard	18.9	11.1
M.I.T.	21.7	0.0
Wisconsin	11.9	23.9
Total	19.1	13.3

TABLE 5

TYPE OF EMPLOYERS OF
DOCTORATES IN 1969
(PERCENTAGE DISTRIBUTION)

Type of Employer	Berkeley	Chicago	Columbia	Harvard	M.I.T.	Wisconsin	Total
Academic	86.4	74.2	68.8	59.1	58.6	63.9	66.8
Business	6.8	7.2	10.7	13.4	24.1	7.3	11.0
Government	3.4	11.3	5.4	18.8	10.3	12.1	12.3
Nonprofit	1.7	3.1	6.4	3.4	3.4	3.6	3.7
All other	1.7	4.1	8.6	5.2	3.4	13.2	6.3
Total known	100.0	99.9	99.9	99.9	99.8	100.1	100.1
Unknown*	27.5	8.5	19.8	7.6	3.3	42.0	18.7

* Percent of all doctorates.

source suggests that those not belonging to the AEA are much more likely to be in business, government, and other professions or deceased or retired, and less likely to be in academic work.[4] The information on doctorates was especially incomplete for Wisconsin and Berkeley, and appreciably so for Columbia. For Columbia, we can employ the occupational distribution from alumni records to distribute the unknown group; this procedure yields a slightly lower academic percentage, about

---

[4] The differences are illustrated by the percentage distribution by occupation from the two sources:

	AEA *Handbook*	Alumni Records
Academic	74.8	44.0
Business	10.2	16.0
Government	10.5	23.0
Nonprofit bodies	3.8	4.0
Other professions	0.0	5.0
Retired or deceased	0.7	8.0

TABLE 6

LEADING ACADEMIC AFFILIATIONS OF DOCTORATES IN 1969 BY INDIVIDUAL SCHOOL

1969 Affiliation	Berkeley	Chicago	Columbia	Harvard	M.I.T.	Wisconsin	Total
Berkeley	3	1	2	2	1	...	9
Chicago	...	6	...	...	...	...	6
Columbia	...	2	6	2	1	...	11
Harvard	1	...	1	7	...	...	9
M.I.T.	...	1	2	2	2	...	7
Wisconsin	4	1	2	2	1	4	14
CCNY	1	1	6	1	...	...	9
Univ. of Illinois	1	2	1	2	...	2	8
Univ. of Michigan	2	...	3	2	1	...	8
Northwestern	1	3	...	1	1	1	7
NYU	...	...	3	3	...	1	7
Univ. of Washington	3	2	...	...	...	1	6
Yale	...	...	...	5	1	...	6
Miami Univ.	...	1	...	1	...	3	5
Univ. of Missouri	1	...	...	1	...	3	5
Rutgers	...	...	4	1	...	...	5
Pittsburgh	...	1	1	1	1	...	4
Princeton	...	1	...	3	...	...	4
Stanford	...	1	...	3	...	...	4
UCLA	1	1	...	2	...	...	4
Wayne	...	2	...	1	...	1	4

65. Alumni information is not available for Berkeley or Wisconsin but we infer that with fuller information these schools would have relatively fewer academicians, so Berkeley would approach the other schools and Wisconsin fall below them in this area. Similarly, the share of doctorates of each school in business and government would probably have become less unequal.

The two-thirds of the doctorates who enter academic life are more widely dispersed among schools than they were in their undergraduate days. Nevertheless there is a marked penchant of the schools to hire their own doctorates. In our sample there is only one exception to the rule that the graduate school attended is the largest single employer of its doctorates (see table 6). The explanation for this home recruitment is probably that the information possessed about one's doctorates is much more intensive and reliable than that provided by other schools. This argument applies primarily to young doctorates, but it is also true that an assistant professor at any school has 10 or 20 times as large a probability of promotion at that school as a comparable scholar elsewhere.

The detailed analysis of geographical location of doctorates confirms the belief that there is a certain regionalism in the academic market: 52 percent of the Harvard doctorates were (in 1969) teaching in schools in the northeast; 46 percent of the Berkeley doctorates were in the west; 55 percent of Wisconsin doctorates were in the north central area; etc. A

TABLE 7

JOURNAL PUBLICATIONS OF ALL DOCTORATES FROM SELECTED SCHOOLS, 1950–68

	Berkeley	Chicago	Columbia	Harvard	M.I.T.	Wisconsin
Doctorates, 1950–55 ..	80	106	116	225	30	143
No. publishing at least one article	58	79	82	153	18	68
% publishing at least one article	72.5	74.5	70.7	68.0	60.0	47.6
Total articles per doctorate:						
All doctorates	4.65	5.56	3.66	4.94	4.17	2.06
Doctorates with publications	6.41	7.46	5.18	7.26	6.94	4.32
No. of articles published	372	589	425	1,111	125	294
% of articles in value and monetary theory	14.8	26.3	22.6	18.4	N.A.	N.A.

strikingly similar geographical segmentation held for the baccalaureate origins of the doctorates, so one may assert that there is a surviving element of the doctrine of noncompeting groups.[5]

Publication by the Doctorates

The total number of doctorates conferred by the schools during 1950–55 is reported, together with their aggregate journal publications during 1950–68, in table 7. The enumeration of publications is a trifle chauvinistic because we are limited to English-language journals, but apparently correction for this distortion would be minor.[6]

The fraction of doctorates who publish is large in the major schools: about two-thirds publish at least one article. The articles per doctorate

---

[5] The full tabulation is

PH.D. INSTITUTION	REGION	PERCENTAGE OF DOCTORATES WHO WERE LOCATED IN REGION AS	
		Undergraduates	College Teachers
Berkeley	West	53.2	46.0
Chicago	North central	56.1	41.5
Columbia	North east	61.5	62.7
Harvard	North east	49.4	52.4
M.I.T.	North east	42.9	46.2
Wisconsin	North central	61.7	55.1

[6] Columbia gave 13 doctorates to students with foreign-language degrees, but eight of the 13 remained permanently in the United States. The fraction of foreign-language degree holders was similar in M.I.T. and Wisconsin (12%–14%) and somewhat smaller in Berkeley, Chicago, and Harvard (4%–8%).

TABLE 8

DOCTORATES BY TOTAL ARTICLES PUBLISHED ON ALL ECONOMIC SUBJECTS, 1950–68

ARTICLES	DOCTORATES						
	Berkeley	Chicago	Columbia	Harvard	M.I.T.	Wisconsin	Total
0	22	27	34	72	12	75	242
1	10	14	24	26	1	23	98
2	5	4	10	20	5	10	54
3	2	14	12	15	0	9	52
4	9	6	5	15	0	5	40
5	7	6	4	11	0	4	32
6–10	16	17	15	28	8	10	94
11–15	4	9	10	22	3	5	53
16–20	3	3	0	8	1	1	16
21 and over*	2	6	2	8	0	1	19
Total	80	106	116	225	30	143	700

* Doctorates with maximum number of articles published are as follows: Berkeley, K. Fox, 22; Chicago, E. Mishan, 41; Columbia, K. Arrow, 36; Harvard, R. Solow, 44; M.I.T., G. Strauss, 18; Wisconsin, R. Lampman, 26.

are highest for Chicago graduates, nearly equal for Berkeley, M.I.T., and Harvard graduates, and at a much lower level for Wisconsin doctorates—only the last difference is statistically significant.

The inequality among doctorates is immense (tables 8, 9). In the fields of value and monetary theory, half of those who published published only one article; the corresponding figure for articles on all subjects is 21 percent. The most prolific two doctorates wrote 13.8 percent of the articles in value theory, and the top 15 (of a total of 107) wrote 45.8 percent of the articles; the corresponding percentages in monetary and fiscal theory were 10.6 and 46.3 (and a total of 96 doctorates). If egalitarianism ever reaches the domain of scholarly research, the journals will fall into evil days.

Our study covers approximately a half-life of the doctorates—an average of 16 years of publication. Writing and not writing are both strongly habit forming, so the skewness of the profile of publication will increase with time.

The skewness of the distribution of articles of the doctorates raises a troublesome question in measuring attributes of the citations of the doctorates. Should we reckon Mishan the equal of 17 less prolific Chicago doctorates, or Arrow the equal of 15 less prolific Columbia graduates? We shall generally employ a method of counting citations (described below) which does not allow multiple citations from one article but allows each article an equal voice. Some comparisons with the one-man one-vote methods of measurement will also be made.

TABLE 9

DOCTORATES BY ARTICLES PUBLISHED IN VALUE AND MONETARY THEORY, 1950–68

	DOCTORATES						
ARTICLES	Berkeley	Chicago	Columbia	Harvard	M.I.T.	Wisconsin	Total
Value theory:							
1	8	13	8	16	4	...	49
2	5	3	5	6	3	...	22
3	...	3	3	3	...	...	9
4	1	2	1	4	...	...	8
5	2	1	...	1	...	...	4
6–10	...	3	1	6	1	...	11
11–15	...	...	...	2	...	...	2
16–20	...	1*	...	...	...	...	1
21 and over	...	...	1†	...	...	...	1
Total	16	26	19	38	8		107
Monetary and fiscal theory:							
1	9	10	8	20	...	7	54
2	3	6	4	2	...	...	15
3	1	4	1	4	...	...	10
4	...	3	1	...	...	...	4
5	1	1	...	2	...	...	4
6–10	...	3	...	4	...	...	7
11–15	...	...	1	...	...	1	2
Total	14	27	15	32		8	96

* Mishan, 20 articles.
† Arrow, 24 articles.

II. The Citation Approach

The citation practices by and toward a scholar are becoming a popular source of information on his intellectual debtors and creditors. Citations are of course a fallible index for any one person: styles of citation vary enormously. The erudite scholar (rightly or wrongly associated with an older Germanic tradition) who displays his learning in his footnotes is hardly recording the strong intellectual influences which have acted upon him. The ostensibly casual scholar (surely trained at Oxbridge) considers citation beyond a name, preferably misspelled, to be a pedantical display. The scholars of all schools are united in their penchant for citing themselves. Some men are careful not to cite their greatest debts. All such differences, one is entitled to believe, are much reduced in magnitude when we combine the citation practice of a substantial number of scholars, as we shall do.

To say that individual idiosyncrasies are submerged in a statistical aggregate is not to say that the aggregate is a correct measure. The nature of intellectual influence is most varied. The direction and, perhaps, the extent of influence are reasonably clear when we follow Friedman and employ permanent and transitory income concepts in a study, or invoke Samuelson on revealed preference. An innovator's work is accepted and used by others. The influence may be most powerful when we simply do not cite at all, and Marshall's theory of long- and short-run equilibrium prices is a fine example. Economists constantly use this distinction, often unaware not only that Marshall introduced it into economics but also that its empirical significance has not been established by Marshall or anyone else. Citation analysis probably works best for fairly recent work which has not had time to be fully absorbed within the literature.

We do not wish to exaggerate the possible weaknesses of citations as a measure of influence. Controversy attracts attention and hence citations, and attention influences scholars. Citations are an easy way to transfer the exposition of a theory or problem from your paper to someone's else, so in the larger view citations reveal a form of intellectual collaboration. To some degree citations *are* influence, for they influence the reading by readers of the citing paper.

The Domain of Citations

Two large areas of economic literature were chosen for the citation analysis: "value" theory (as we term it) and monetary and fiscal theory. In terms of the categories of the *Index of Economic Articles*, the areas covered are:

I. Value Theory
    2.1 Value, Price and Allocation Theory
    2.2 Factors of Production and Distributive Shares
    15.23 Monopoly. Concentration. Competition
II. Monetary and Fiscal Theory
    2.3 Aggregative and Monetary Theory
        (Specifically 2.30–2.33)
    9.6 Prices. Inflation. Deflation
    9.9 Monetary Policy
        (Specifically, 9.90–9.93)
    10.2 Fiscal Policy for Economic Stabilization and Growth
    10.5 Public Debt and Debt Policy
    13.3 War Finance and Stabilization Policies
        (Specifically, 13.32–13.33)

For each article (in the indexed journals) written between 1950 and 1968 by those receiving doctorates between 1950 and 1955 the number of times each economist is cited and the type of citation (favorable, unfavorable; or neutral) has been compiled. The publication period has been divided into three periods (1950–57, 1958–62, and 1963–68) in order to study the changes in citation practices—usually without significant result, perhaps because of small sample sizes. It is on this body of evidence on influence and attitude that our main study rests.

How Should Citations Be Counted?

The literal number of times a man is mentioned is one possible measure, in a given article, of the extent of his influence. It is a lax measure, since a man's name may be cited a dozen times in one paragraph on one point and then dropped. In our count of total citations we have counted all citations subject to the limitation of one citation of any authority per paragraph.

The total number of citations, however, is not necessarily the proper measure of a person's influence. The leading doctoral citers have strong enthusiasms, so they heap large hills of citations upon favored (or disfavored!) fellow economists. That compulsive citer, E. J. Mishan, cites Hicks 66 times, Little 55 times, and Kaldor 43 times; Myron Gordon thinks often of Modigliani (55 citations); and similar things can be said of Robert Dorfman (Leontief, 41 times), Robert Strotz on Von Neumann and Morgenstern (each 31 times), etc. Some indication of this concentration of attention is given by the tabulation of leading doctorate citers in value theory in reference table 1.[7]

These bursts of enthusiasm are perhaps a source only of mild surprise, but they raise a question: What kind of influence do we wish our citations to measure? We shall find that Walras is the third most frequently cited name in value theory with 160 citations, of which 93 came from Kuenne (and only 10 other of the 30 most prolific doctorate citers even mention Walras). That seems a much narrower impact than that of Alfred Marshall, who was cited only 107 times but who received only 18 citations from the most frequent citers (Mishan and H. Levin) and was cited by 17 of the 30 leading doctorate citers. There are numerous other examples of narrow and of wide circles of citation. Consider the sample from citations in value theory shown in table 10.

There are various ways in which we can choose our citation count in order to measure better the breadth of influence of the authorities. The

---

[7] A set of the following reference tables may be obtained by writing to the authors: (1) Most Frequent Doctorate-Citers, Economic Theory, 1950–68; (2) Number of Citations in Articles on Value Theory to the Most Frequently Cited Authorities, 1950–68; (3) Number of Citations in Articles on Monetary and Fiscal Theory to the Most Frequently Cited Authorities, 1950–68; (4) Number of Articles in Price Theory Published by the Leading Authorities; (5) Number of Articles in Monetary and Fiscal Theory Published by the Leading Authorities; (6) Citations of Authorities in Price Theory by Doctorates and Authorities: Number of Articles in Which Cited (at Least Once); (7) Citations of Authorities in Monetary and Fiscal Theory by Doctorates and Authorities: Number of Articles in Which Cited (at Least Once).

TABLE 10

CITATIONS IN VALUE THEORY

Name	Total Citations	Citations from Three Most Frequent Citers of Each (%)
Hicks	268	50.4
Samuelson	209	39.7
Walras	160	80.0
Little	151	73.5
Marshall	107	44.9
Kaldor	83	79.5
Von Neumann	68	83.8
Stigler	67	31.1
Lange	63	79.4
Knight	49	38.8

count of number of doctorates ever citing an authority, already employed previously, seems too severe. On the other hand, if Kenneth Arrow repeatedly cites Samuelson (40 times) in various articles, that ought to count for more than his single citations of Bain and Marschak. A fully rational weighting of citations would be based upon an acceptable measure of intellectual influence, which we utterly lack.

As a compromise among the available measures, we shall use as our fundamental measure of citations the number of articles in which an authority is cited at least once. For the purpose of our primary interest, the measurement of the impact of institutional choice upon doctorates, it is desirable to have a measure of impact which does not allow a few heavy citers to overwhelm the results. The use of "at least one" citations compromises by eliminating multiple references to an authority within one article and yet allows the more prolific doctorates a louder voice, which we believe they should have, in revealing intellectual influence.[8] When we say (table 11) that Samuelson was cited 72 times, we mean that there were 72 articles written by the doctorates in which he was cited at least once.

Favorable and Unfavorable Citations

The two graduate students who performed the citation analysis were requested to classify the citations as "favorable," "unfavorable," and "neutral"—using the last category if there was serious doubt whether a citation was favorable or unfavorable. Of the 5,581 citations they

[8] The count of names cited at least once is much more accurate than that of total citations. In experimental replications we found that an eligible authority was seldom omitted, but the count of citations was systematically too low by roughly 4%–5%. The top 30 names cited at least once in value theory and the top 30 names in total citations in value theory, a combined total of 35 names, has a rank correlation of .76.

classified, some 648 were favorable and 566 unfavorable—mild contradiction of Hugh Dalton's famous remark that economists take in each other's dirty laundry to wash.

Usually a citation is easy to classify. For example, Eckstein (with Wilson) is surely being favorable toward H. M. Levinson when he states: "We are considerably indebted to the pioneer work of Levinson, who emphasized the importance of wage rounds and key bargains in his qualitative account. Levinson discovered the central significance of profits from his analysis of annual data for two-digit manufacturing industries. He also made the important negative discoveries that productivity and output changes were not significant variables in the explanation of wages." (Eckstein and Wilson 1962, p. 402).[9] Often, however, the classification taxes one's casuistical skill. Consider Hicks's review article on Patinkin, which begins by stating: "The main things I have learned from [the book] are not what the author meant to teach me" (1957, p. 278). It is not complimentary to be told that one did not understand his own message; it is complimentary for an economist to be able to teach Sir John anything—so the passage is classified as neutral. Although there is no unambiguously correct classification, we believe the classification has been fairly consistent.[10]

Who Is Cited?

It would be possible to tabulate, in our universe of articles written by the doctorates in economics between 1950 and 1968, the frequency with which every economist (or other person) is cited. For the purpose of determining academic loyalties and intellectual affiliations, this list of potential authorities was curtailed in the interest of economy in tabulating the basic data. Accordingly, we tabulate only citations to publications of (1) members of the 1950–55 faculties of the schools under study and (2) leading figures in the literature.[11]

The selection of authorities was narrower than, in retrospect, we wish it had been. The nonfaculty authorities were compiled from fairly brief surveys of the appropriate literature. Thus Don Patinkin, one of the most cited of monetary economists, was not in the list of eligible price theorists by our criterion, so citations to him were not tabulated in articles on price theory. It follows that we have excluded many citations, and in particular those to many of the faculty at schools whose doctorates are not included in our study. The omitted names, to be sure, would have appeared chiefly in the bottom halves of the distributions.

---

[9] An unequivocally unfavorable citation may be illustrated by S. C. Tsiang (1956, p. 540) on Lerner: "Abba Lerner's proof of the equivalence of the liquidity preference and loanable funds theories was achieved only at the expense of total distortion of the latter. For to establish the identity of the two theories, he interpreted the two main components of the demand and supply of loanable funds, viz., savings and investment, in the *ex post* sense in which they are always identical. This made nonsense of the loanable funds theory. . . ."

[10] In a small replication of the classification of citations, the differences, which ran around 7 percent, were systematic. We consistently classified more citations as neutral and fewer as favorable than did Glen Gilchrist and Richard Ippolito, to whom we are indebted for the analyses of the citations.

[11] The leading figures were identified by the frequency of citation in Ellis (1948), Haley (1952), and *Surveys of Economic Theory* (1965 and 1966).

The citations by doctorates are of course not typical of the journal literature. A given age-cohort will be more homogeneous in its interests, reading, and writing than the profession at large, and this homogeneity is strengthened by the restriction to a few major schools. However, these schools produce a large share of the leading economists, and it could be argued that the citation practices of doctorates of these schools during their prime years of scientific work provide better measures of influence than a broader group would.

The citation counts are also much influenced by the immense concentration of writings and of citations in a few unusually prolific doctorates. In this respect the doctorates are not very different from the profession at large: most economists write little, a few write much; and of the writers, some cite parsimoniously and others as if every name will yield a job offer.

The frequencies of citation we shall report in the next section will inevitably invite interpretation as measures of the importance or influence of the cited authorities. This is in good part an inescapable inference. In monetary and fiscal economic theory we report as the leaders Keynes, 61 citations; Friedman, 45; Hicks, 34; Patinkin, 31; and Pigou, 31. No one can dispute that the first four of these were among the most influential intellectual leaders in this subject area in the postwar decades. (Pigou probably owed his popularity to discussions of the Pigou effect.) Yet our measures may be quite inaccurate when we drop farther down the lists. An economist, you must be reminded, was not counted among the citees, no matter how often he was cited, unless he was mentioned several times in the survey articles or was on the faculty of one of the schools included in our study (five schools in each area, with four appearing in both lists).[12]

III. The Influence of Schools on Doctorates

We come at last to the citations. Let us present at once the two samples of articles which we have analyzed: those dealing with value theory and with monetary and fiscal economics. These lists of leading "authorities" (this good Victorian word must not lead us to forget that some were, to those citing them, adversaries or dunces) are presented in tables 11 and 12 on the basis of "at least one" citation per article, and the corresponding total citation lists are presented in reference tables 2 and 3. More economists are cited in articles on value theory than in those on monetary theory, so we carry the listing of names farther in the former field.

---

[12] The following are illustrative names omitted from our master lists in both value theory and monetary economics: Böhm-Bawerk, Eisner, Hawtrey, Houthakker, Jorgenson, Lipsey, Tinbergen, and Wicksell. In addition, Otto Eckstein and Edwin Kuh were inadvertently omitted from the monetary theory list.

We emphasize for the last time that many distinguished names are omitted from the lists because of the nature of the criteria for inclusion. Even with this large proviso, it is safe to call the following the leaders of general economic theory so far as the doctorates are concerned:[13] Hicks, 105 citations in the two tables; Keynes, 91; Samuelson, 90; Friedman, 78; Pigou, 56. (The first four names would be the same if we used total citations; the fifth would be Patinkin.) If we had broadened our canvass to the other fields of economics, many names would have risen in relative frequency: Gardiner Means probably would have dominated industrial organization; Johnson, Metzler, Meade, Nurkse, and Viner, international trade; Slichter and Douglas, labor economics; and so on. If we had broadened the canvass to other schools, the effects are more problematical, but probably the ranking of names near the tops of the lists would be little changed.

There is enough resemblance between the lists of authorities of the various schools to suggest that there is indeed only one theory of value and one theory of monetary and fiscal economics. Thus the doctorates of each of the three large schools (Chicago, Columbia, and Harvard) cited each of the 26 leading authorities on value theory and each of the 18 leading monetary and fiscal theorists. As a more summary measure of the concordance among the schools, we tabulate the correlation coefficients (calculated from log $[C + 1]$ where C is the citation count) in table 13.[14]

In value theory the various correlation coefficients (for the 49 leading authorities) generally fall within the range of .3 and .4, with three conspicuous exceptions: Berkeley and Columbia are uncorrelated, and Chicago and Columbia are closely correlated. In monetary and fiscal economics the correlations are lower—the first of the evidences we shall find of a lesser agreement on authorities in this branch of economics.[15]

---

[13] For evidence that citations measure the *quality* of scientific output, see Cole and Cole (1973, chap. 2). The evidence seems much weaker to us than to the Coles.

[14] The log transformation reduces the correlations but yields much more plausible scatter diagrams. With actual citations the regression line goes through the origin and has a slope determined by citations to a few leading authorities.

[15] In interpreting these correlation coefficients it is desirable to note a fairly high measure of instability in the rankings. When we correlated citations of one half of the doctorates against the other half (the division being random), the correlation coefficients for 49 authorities in value theory were Berkeley .56, Chicago .28, Columbia .27, and Harvard .07!

TABLE 11

Articles on Value Theory in Which Authorities Were Cited at Least Once, 1950–68

Authority	Ph.D.'s from					Five-School Total
	Berkeley	Chicago	Columbia	Harvard	M.I.T.	
P. A. Samuelson	6	15	17	27	7	72
J. R. Hicks	9	18	13	25	6	71
A. Marshall	4	16	10	14	2	46
G. J. Stigler	1	15	14	9	1	40
K. J. Arrow	0	11	5	15	2	33
M. Friedman	1	9	12	11	0	33
W. J. Baumol	4	5	6	16	1	32
J. M. Keynes	3	7	4	13	3	30
L. Walras	6	5	7	9	1	28
T. Scitovsky	3	10	5	7	2	27
F. H. Knight	5	9	7	5	0	26
F. Modigliani	0	6	3	16	1	26
V. Pareto	1	7	10	7	1	26
A. C. Pigou	2	10	5	6	2	25
W. Leontief	4	4	1	13	1	23
J. von Neumann	1	5	9	8	0	23
O. Morgenstern	0	5	9	8	0	22
R. M. Solow	3	5	4	8	1	21
N. Kaldor	2	8	2	8	0	20
O. Lange	0	3	8	7	2	20
K. E. Boulding	2	7	3	7	0	19
I. M. D. Little	0	10	5	3	1	19
H. Hotelling	0	5	5	7	0	17
H. A. Simon	0	3	6	8	0	17
A. P. Lerner	1	5	4	6	0	16
J. Robinson	5	2	1	7	1	16
J. S. Bain	0	0	8	6	1	15
T. Koopmans	2	4	5	3	1	15
J. Marschak	3	2	2	8	0	15
J. A. Schumpeter	2	1	3	8	1	15
A. Bergson	0	2	7	5	0	14
E. H. Chamberlin	2	2	2	8	0	14
J. Duesenberry	3	2	3	5	1	14
M. W. Reder	0	3	4	7	0	14
R. G. D. Allen	1	4	2	4	1	12
A. Cournot	1	1	5	4	1	12
R. Dorfman	1	1	0	9	0	11
R. F. Harrod	0	2	2	6	1	11
L. R. Klein	2	0	1	8	0	11
E. Kuh	0	1	1	9	0	11
J. Tobin	0	2	2	7	0	11
G. S. Becker	0	1	2	5	2	10
J. M. Clark	1	1	4	4	0	10
P. H. Douglas	2	2	0	6	0	10
J. T. Dunlop	0	1	1	8	0	10
O. Eckstein	0	2	1	7	0	10
W. J. Fellner	1	0	0	8	1	10
L. A. Metzler	0	2	5	3	0	10
D. H. Robertson	3	4	0	3	0	10
Other persons cited:						
5 or more times (39 persons)	22	52	57	115	9	255
1–4 times (75 persons)	18	28	32	70	4	152
Total (163 persons)	127	325	324	595	59	1,430

TABLE 12

ARTICLES ON MONETARY AND FISCAL THEORY IN WHICH AUTHORITIES WERE CITED AT LEAST ONCE, 1950–68

AUTHORITY	PH.D.'S FROM					FIVE-SCHOOL TOTAL
	Berkeley	Chicago	Columbia	Harvard	Wisconsin	
J. M. Keynes	6	26	11	17	1	61
M. Friedman	1	21	14	9	0	45
J. R. Hicks	4	13	5	11	1	34
D. Patinkin	3	18	5	5	0	31
A. C. Pigou	1	18	6	6	0	31
A. H. Hansen	3	7	5	8	2	25
J. Tobin	1	8	4	8	3	24
A. P. Lerner	4	7	6	3	1	21
G. Haberler	3	1	5	11	0	20
A. Marshall	3	7	3	5	0	18
P. A. Samuelson	3	5	2	5	3	18
E. S. Shaw	3	2	5	7	1	18
J. Duesenberry	1	4	3	8	1	17
R. M. Goodwin	2	2	1	12	0	17
J. Gurley	2	2	2	9	2	17
W. L. Smith	0	4	5	7	1	17
L. A. Metzler	2	8	1	5	0	16
L. R. Klein	3	7	2	2	0	14
D. H. Robertson	3	5	0	5	0	13
R. M. Solow	1	4	3	5	0	13
W. J. Baumol	1	0	2	8	1	12
M. Bronfenbrenner	2	0	4	4	2	12
J. M. Clark	3	3	0	6	0	12
A. Smithies	0	3	2	6	1	12
A. F. Burns	3	1	1	4	1	10
E. D. Domar	2	3	0	5	0	10
I. Fisher	0	7	1	2	0	10
A. G. Hart	0	2	4	3	1	10
S. C. Tsiang	2	1	1	6	0	10
R. F. Harrod	2	1	1	5	0	9
H. G. Johnson	1	3	1	4	0	9
J. A. Schumpeter	2	1	1	4	1	9
K. Brunner	0	4	3	1	0	8
S. Kuznets	0	3	0	5	0	8
O. Lange	0	3	1	3	1	8
R. T. Selden	2	1	5	0	0	8
D. G. Johnson	3	3	1	0	0	7
F. H. Knight	1	4	0	2	0	7
T. Koopmans	0	3	0	4	0	7
E. R. Rolph	1	2	2	2	0	7
S. H. Slichter	0	2	1	4	0	7
W. J. Fellner	4	0	1	1	0	6
W. Leontief	1	2	1	2	0	6
A. W. Marget	1	2	2	1	0	6
A. E. Rees	0	2	0	4	0	6
R. Turvey	0	1	1	4	0	6
J. W. Angell	1	1	2	1	0	5
O. Brownlee	0	4	1	0	0	5
J. T. Dunlop	2	1	0	2	0	5
J. K. Galbraith	1	0	1	3	0	5
E. J. Hamilton	1	4	0	0	0	5
F. Machlup	0	0	0	5	0	5
L. W. Mints	0	3	1	1	0	5
A. W. Phillips	0	1	0	4	0	5
Other persons cited 1–4 times (56 persons)	16	36	19	40	3	114
Total (110 persons)	101	276	148	294	27	846

TABLE 13

Correlation between Schools in Citation of Authorities: Articles in Which Cited (at Least Once)

	Chicago	Columbia	Harvard	M.I.T.	Wisconsin
Value theory, leading 49 authorities:					
Berkeley	.36	.07	.33	.39	...
Chicago	...	.60	.39	.42	...
Columbia	...	...	.27	.39	...
Harvard	...	...	...	.47	...
Monetary and fiscal theory, leading 54 authorities:					
Berkeley	.19	.31	.23	...	.21
Chicago	...	.46	.25	...	.09
Columbia	...	...	.31	...	.42
Harvard	...	...	...	...	.40

Note.—Correlations between log $(C_i + 1)$ and log $(C_j + 1)$, where C_i is the number of citations of doctorates from school i.

Aside from the penchant of doctorates in citing their own faculty, to which we shall shortly turn, it is difficult to find systematic differences among the schools in citation practices. In value theory, it is possible to distinguish a set of economists who make relatively large use of mathematics: the doctorates of all schools cited them with about equal relative frequency. Again, it is possible to distinguish a set of economists who are relatively active in welfare economics: here the differences were larger, but only the doctorates from M.I.T. showed a morbid interest in the subject. Finally, if one distinguishes economists who have done extensive empirical work, they are about equally frequently cited by doctorates of each school—possibly somewhat less by those from Berkeley and M.I.T. In general this finding is in keeping with the impression held by many economists (including ourselves) that value theory is a reasonably well-defined body of knowledge which most competent economists accept.

In the monetary and fiscal area the differences among schools are larger. If we loosely classify a group as Keynesians,[16] there is no important difference among the doctorates of the various schools—about one-fifth of each school's citations are to this group (indeed Chicago is highest with 22.5 percent). But a similar grouping of monetarists[17] reveals large differences: one-tenth of Harvard and Berkeley doctorates cite the monetarists against one-sixth for Columbia and one-fifth for Chicago.

Parochial Loyalty

The doctorates of a school naturally cite relatively frequently the publications of the faculty of that school. Familiarity alone would lead to

---

[16] Keynes, Hansen, Tobin, Lerner, Samuelson, Goodwin, Klein.

[17] Friedman, Patinkin, Robertson, Tsiang, Brunner, Marget, Hamilton, Mints.

TABLE 14

"AT LEAST ONE" CITATIONS IN VALUE THEORY
(PERCENTAGE DISTRIBUTION)

Doctorates	Authority on Own Faculty	Authority on Faculty of Other Four Schools	Authority Elsewhere
Berkeley	7.1	41.7	51.2
Chicago	13.2	26.8	60.0
Columbia	9.6	32.1	58.3
Harvard	13.7	31.1	55.2
M.I.T.	15.3	27.1	57.6

this result and it is reinforced by friendships and possibly by indoctrination. The same factor of familiarity—greater knowledge—leads to a preference of English economists for citing Englishmen, and similarly for Americans, as we shall see later. Knowledge probably also leads to the relatively heavy concentrations of the citations of the doctorates on the faculties of the other leading schools. To give one sample from value theory, the distribution of "at least one" citations is shown in table 14. Of course our universe of authorities excludes many economists, all outside these schools, but nevertheless the concentration of citations on the faculties of the largest producers of doctorates is substantial.

We present the cross-tabulation of doctorates by faculties in tables 15 and 16. An index of "parochialism" has been calculated as follows: (*a*) Calculate the percentage distribution of school *i's* citations in the publications of the three other large schools, excluding citation by the doctorates of their own faculty.[18] (*b*) Average these unweighted ("expected") distributions of citations for the four large schools, and adjust to 100 percent. These are the "expected" relative frequencies of citations of the faculties of the five schools.[19] (*c*) Divide this expected percentage of citations into the actual percentage for that school's doctorates. This procedure yields an index of parochialism. For example, in value theory, Berkeley doctorates cited ("at least once") their own faculty 14.52 percent of the total citations to the five faculties. The average Berkeley faculty share in the citations of the doctorates from the other large schools was 19.20 percent, so the index (14.52/19.20 = 0.76) shows *negative* parochialism in this category. The indexes of parochialism are a trifle parochial because they rely on the doctorates' citations of faculty of other schools, but it would require a complex study of citations (perhaps by leading scholars) to improve upon them.[20] The "at least one" indexes of parochialism, together with the patterns of citations, are also given in tables 15 and 16.

---

[18] The school with the smallest number of articles is omitted.

[19] The larger schools are unweighted to reduce the influence of the largest schools.

[20] See the discussion of citations by authorities, below.

TABLE 15

"AT LEAST ONE" CITATIONS AMONG FACULTIES OF FIVE SCHOOLS: VALUE THEORY (PERCENTAGE DISTRIBUTION)

SCHOOL OF DOCTORATES	FACULTY CITED					
	Berkeley	Chicago	Columbia	Harvard	M.I.T.	Total
Berkeley	14.5	32.3	8.1	27.4	17.7	100.0
Chicago	13.8	33.1	13.8	20.8	18.5	100.0
Columbia	17.0	29.6	23.0	12.6	17.8	100.0
Harvard	20.4	19.6	10.7	30.4	18.9	100.0
M.I.T.	12.0	16.0	8.0	28.0	36.0	100.0
Indexes of parochialism	.76	1.19	1.89**	1.44**	1.83*	

* Significantly different from unity at 10% level.
** Significantly different from unity at 5% level.

TABLE 16

"AT LEAST ONE" CITATIONS AMONG FACULTIES OF FIVE SCHOOLS: MONETARY AND FISCAL THEORY (PERCENTAGE DISTRIBUTION)

SCHOOL OF DOCTORATES	FACULTY CITED					
	Berkeley	Chicago	Columbia	Harvard	Wisconsin	Total
Berkeley	23.1	25.0	17.3	28.8	5.8	100.0
Chicago	3.6	61.8	10.0	23.6	0.9	100.0
Columbia	16.2	29.4	17.6	30.9	5.9	100.0
Harvard	9.4	24.5	10.8	51.8	3.6	100.0
Wisconsin	7.7	15.4	15.4	38.5	23.1	100.0
Index of parochialism	1.87*	2.05**	0.98	1.49**	4.95**	

* Significantly different from unity at 10% level.
** Significantly different from unity at 5% level.

All of the statistically reliable indexes of parochialism are above unity.[21] Harvard is consistently parochial, and Chicago (money) and Columbia (value) also strongly so. The indexes are generally high in the schools with few articles by doctorates (M.I.T. and Wisconsin). If one analyzes total citations—the tables are based upon "at least one" citations—the evidence of parochialism is usually even stronger.

When the period is divided into three subperiods, the parochial indexes become less stable but show no clear tendency to decline with the passage of time.

Parochial indexes were also calculated for doctorates publishing (1)

---

[21] The tests of differences of proportions from unity are not exact because it is not possible to specify the precise number of degrees of freedom: the ratios of the other three schools (whose unweighted average we use) is based upon differing numbers of citations.

TABLE 17

CLASSIFICATION OF CITATIONS
(PERCENTAGE DISTRIBUTION)

	Five Faculties	Other Authorities	Total
Value theory:			
Favorable	9.4	4.8	6.7
Neutral	84.6	86.6	85.8
Unfavorable	6.0	8.6	7.6
Monetary and fiscal theory:			
Favorable	24.1	18.2	20.6
Neutral	60.3	67.5	64.5
Unfavorable	15.6	14.3	14.8

three or less articles and (2) four or more articles. It is a plausible conjecture that the more productive scholars will be less parochial because they are more extensively engaged in research—and less dependent upon their doctorate institution for academic support. No such difference was found.

The classification of citations as favorable or unfavorable (or neutral) yields further information on the intellectual associations among schools.[22] In value theory, the favorable to unfavorable citations run three to two for faculties of the five schools, about even for all authorities, and most references are neutral; opinions are much stronger in monetary and fiscal economics (see table 17). The much higher fraction of nonneutral citations is additional evidence of the controversial nature of that literature. The detailed tabulations are given in table 18.

If we accept the observed ratio of roughly three favorable to two unfavorable references as the standard, there are few important differences among the schools in value theory. Chicago doctorates are unfavorable in their references to Harvard (an attitude which Chamberlin frequently credited to Knight's influence),[23] and similarly M.I.T. doctorates are unfavorable to Chicago, but the general pattern is one of random fluctuation. In monetary and fiscal theory, as usual, the story changes. There is evident parochialism (Harvard doctorates are favorable to their faculty in 44 percent of the citations, unfavorable in 6 percent; and Berkeley is as extreme). Ideological battles are also hinted at by cross-citations: on average, Berkeley and Columbia doctorates' citations of Chicago faculty are unfavorable one-fifth to one-third of the time.

---

[22] When analyzing attitude in citations, we necessarily use total citations, not the at-least-one citation base used elsewhere.

[23] As an early skeptic of monopolistic competition, George Stigler can testify that Knight devoted even less time than compliments to Chamberlin so students had to read the book on their own.

TABLE 18

FAVORABLE AND UNFAVORABLE REFERENCES IN CITATIONS BY DOCTORATES*
(PERCENTAGE DISTRIBUTION)

SCHOOL OF DOCTORATES	FACULTY CITED									
	Berkeley		Chicago		Columbia		Harvard		M.I.T.	
	Favorable	Unfavorable	Favorable	Unfavorable	Favorable	Unfavorable	Favorable	Unfavorable	Favorable	Unfavorable
Value theory:										
Berkeley	20.0	5.0	7.5	9.4	14.3	0.0	10.4	0.0	4.0	4.0
Chicago	7.1	7.1	10.9	7.3	5.6	11.1	2.0	18.0	10.6	2.1
Columbia	3.3	11.5	4.3	8.7	7.8	4.7	2.9	0.0	9.8	0.0
Harvard	16.4	7.8	8.0	6.8	8.5	2.1	10.6	4.0	15.2	2.2
M.I.T.	10.0	0.0	10.0	35.0	0.0	0.0	0.0	0.0	3.0	6.1
Monetary and fiscal theory:										
Berkeley	31.6	0.0	12.7	34.5	12.5	6.2	22.7	0.0	...	...
Chicago	20.0	0.0	22.6	14.5	14.3	14.3	15.6	23.4	...	...
Columbia	23.1	0.0	13.8	22.5	6.2	0.0	24.0	4.0	...	...
Harvard	38.7	32.2	27.0	23.0	13.2	21.0	44.2	6.5	...	...

---

* Wisconsin excluded because of sample size.

IV. The Authorities and the Doctorates

The leading authorities cited in our studies were themselves examined in some detail. A preliminary question which almost asks itself is: Does one become an authority by writing a good deal, or is citation largely independent of writing? The omission of books from our analysis means that our answer must be somewhat partial and biased, but even a partial answer is not without interest. The numbers of articles in value and monetary theory written from 1886 to 1968 are given for each authority in reference tables 4 and 5.[24]

We expect, and observe, some correspondence between publications and citations even though the restriction of our list to leading authorities obscures the influence of publication on citations. Robert Merton has discussed the "Matthew Effect" ("For unto every one that hath shall be given . . . ") in science and given it the interpretation that larger increments of recognition are given a piece of work, the more famous the author already is (Merton 1973, pp. 439–59). As he points out, the fame of the author is a valuable guide to a prospective reader in choosing from the flood of literature that which he will study. It is *not* evident that the acceptance, in contrast to the knowledge, of a piece of work is influenced by its authorship. In any event, the most cited authorities are prolific authorities: the number of articles on price theory published by the five leading authorities was 112, and in monetary economics the corresponding number was 73.

When we regress the number of "at least one" citations in price theory on price theory articles before and after 1950, we find

$$\begin{aligned}\text{citations} = {} & \underset{(t\,=\,2.55)}{11.8} + \underset{(2.40)}{0.93}\ \text{articles}_{1886-1949} \\ & + \underset{(2.78)}{0.95}\ \text{articles}_{1950-68}\ (R^2 = .327).\end{aligned}$$

The comparable equation for citations in monetary and fiscal economics is

$$\begin{aligned}\text{citations} = {} & \underset{(t\,=\,2.31)}{8.09} + \underset{(3.19)}{1.10}\ \text{articles}_{1886-1949} \\ & + \underset{(1.69)}{0.58}\ \text{articles}_{1950-68}\ (R^2 = .254).\end{aligned}$$

Thus the price theory citations give equal attention to publications before and after 1950 and the monetary citations place a higher weight on earlier publications, which is contrary to one's belief that articles should and do depreciate with time. Presumably, the inclusion only of authorities who are cited frequently after 1950 overcomes the depreciation of articles. The linearity of the relationship between citation and publications has also been examined, and there is a trace of increasing returns to articles. The elasticity of citations with respect to articles exceeds unity after the number of articles reaches 19 in price theory and 11 in monetary theory.[25]

---

[24] We omit a few names (Walras, Marshall, Pareto, von Neumann) where for reasons of language or period or discipline, the count would be meaningless. Schumpeter might have been added to this list.

The individual schools do not differ much in the citations of their doctorates with respect to articles published by authorities. In price theory the correlation coefficients for individual schools range from .431 (Columbia) to .601 (Berkeley); in monetary economics the level is lower and the variation wider (from .189 [Berkeley] to .432 [Chicago]). The lower influence of publications upon citations in monetary and fiscal economics than in price theory is in keeping with our consistent finding of greater variability of relationships in the former field.[26]

Relationships among Authorities

The citations of active authorities (those writing after 1950) on each other were studied with a view to discovering what might be called allies and antagonists in scientific work. The hypothesis we propose is that numerous citations, unless they are heavily unfavorable, reveal allies, especially if they are reciprocated. The relationships between authorities have usually been easy to classify even on this simple hypothesis.

First let us begin with the simple relationships. We deem the following to be clear cases of allies (all in price theory):

1. Samuelson (1)-Solow (2). Let ${}_1C_2$ be citations of (2) by (1). Let F be percent favorable, U percent unfavorable. Then

$$ {}_1C_2 = 20\ (F = 10;\quad U = 0), $$

$$ {}_2C_1 = 19\ (F = 15.8;\ U = 0). $$

---

[25] For the entire period, 1886–1968, the equation in price theory is

$$ \text{citations} = \underset{(4.50)}{26.7} - \underset{(1.96)}{1.54}\ \text{articles} + \underset{(3.30)}{0.073}\ (\text{articles})^2\ (R^2 = .531). $$

In monetary economics, there is no relationship for the entire period, but a strong one on publications before 1950, which follows:

$$ \text{citations} = \underset{(5.94)}{17.0} - \underset{(1.51)}{1.24}\ \text{articles} + \underset{(2.87)}{0.13}\ (\text{articles})^2\ (R^2 = .354), $$

where the articles are those published 1886–1949.

[26] In both areas the correlations are consistently higher for "at least one" citations than for total citations, which reinforces our arguments for the former measure.

2. Arrow (1)-Solow (2):

$${}_1C_2 = 6\ (F = 0;\quad U = 0),$$
$${}_2C_1 = 9\ (F = 22.2;\ U = 0).$$

3. Arrow (1)-Samuelson (2):

$${}_1C_2 = 40\ (F = 10.0;\ U = 0),$$
$${}_2C_1 = 20\ (F = 15;\quad U = 5).$$

Probably one should add two pairs:

4. Hicks (1)-Samuelson (2):

$${}_1C_2 = 8\ (F = 12.5;\ U = 12.5),$$
$${}_2C_1 = 56\ (F = 12.5;\ U = 16.1).$$

5. Bergson (1)-Little (2):

$${}_1C_2 = 9\ (F = 33.3;\ U = 11.1),$$
$${}_2C_1 = 6\ (F = 0;\quad U = 0).$$

The clear instances of antagonistic relationships are more numerous if the following qualify:

1. Samuelson (1)-Lerner (2):

$${}_1C_2 = 16\ (F = 0;\ U = 12.5),$$
$${}_2C_1 = 28\ (F = 0;\ U = 67.9).$$

2. Samuelson (1)-Kaldor (2):

$${}_1C_2 = 23\ (F = 0;\ U = 13.0),$$
$${}_2C_1 = 10\ (F = 0;\ U = 70.0).$$

3. Solow (1)-Kaldor (2):

$${}_1C_2 = 13\ (F = 0;\ U = 7.7),$$
$${}_2C_1 = 5\ (F = 0;\ U = 60).$$

4. Solow (1)-Robinson (2):

$${}_1C_2 = 7\ (F = 0;\ U = 42.9),$$
$${}_2C_1 = 18\ (F = 0;\ U = 66.7).$$

5. Arrow (1)-Bergson (2) (not reciprocal):

$${}_1C_2 = 16\ (F = 6.2;\ U = 0),$$
$${}_2C_1 = 56\ (F = 0;\quad U = 48.2).$$

6. Arrow (1)-Little (2):

$$ {}_1C_2 = 26 \ (F = 11.5;\ U = 61.5), $$
$$ {}_2C_1 = 30 \ (F = 0;\quad U = 50). $$

7. Kaldor (1)-Hicks (2):

$$ {}_1C_2 = 10 \ (F = 0;\quad U = 50), $$
$$ {}_2C_1 = 6 \ (F = 16.7;\ U = 33.3). $$

8. Samuelson (1)-Robinson (2):

$$ {}_1C_2 = 15 \ (F = 0;\ U = 6.7), $$
$$ {}_2C_1 = 7 \ (F = 0;\ U = 42.9). $$

There are too few cross-citations to allow analysis of most other possible pairs.[27]

We emerge with two sets of allies. The Cambridge (U.S.) group of Arrow, Samuelson, and Solow are consistently together and are consistently opposed by the Cambridge (England) group of Kaldor and Robinson. Hicks comes much closer to the American Cambridge. In addition, if we analyze pairs for which there are considerable cross-citations but less than the number (four each way) required above, we can distinguish several additional groups of allies:

1. The monopolistic competition school, including Bain, Chamberlin, Modigliani, and Steiner.[28]

2. An input-output (and related branches) group, containing Dorfman, Leontief, Solow.[29]

3. A vague hint of what others call the Chicago school (although Stigler is classified as Columbia): Becker, Friedman, and Stigler.[30]

---

[27] The Solow (1)-Hicks (2) pairing is essentially neutral (${}_1C_2 = 3$, ${}_2C_1 = 5$, all but one of eight citations neutral); and similarly for Arrow (1)-Baumol (2) (${}_1C_2 = 4$, ${}_2C_1 = 5$, on balance favorable); and Samuelson (1)-Little (2) (${}_1C_2 = 8_2$, $C_1 = 4$, favorable on average).

[28] Letting Bain, Chamberlin, Modigliani, and Steiner have respective numbers 1, 2, 3, and 4:

$$ {}_1C_2 = 9 \ (F = 77.8);\ {}_2C_1 = 1 \ (F), $$
$$ {}_2C_4 = 12 \ (F = 33.3;\ U = 8.3);\ {}_4C_2 = 1 \ (F), $$
$$ {}_3C_1 = 30 \ (F = 33.3;\ U = 13.3). $$

[29] Letting Dorfman = 1, Leontief = 2, and Solow = 3:

$$ {}_1C_2 = 35 \ (F = 8.6;\ U = 5.7), $$
$$ {}_3C_1 = 9 \ (F = 33.3), $$
$$ {}_3C_2 = 20 \ (F = 5;\ U = 5). $$

[30] Letting Becker = 1, Friedman = 2, and Stigler = 3:

$$ {}_1C_3 = 9 \ (F = 22.2), $$
$$ {}_2C_3 = 3 \ (F = 66.7). $$

TABLE 19

CITATIONS TO GROUPS OF ALLIED ECONOMISTS IN PRICE THEORY: FAVORABLE AND UNFAVORABLE (PERCENTAGE DISTRIBUTION)

CITED GROUP	DOCTORATE					
	Berkeley	Chicago	Columbia	Harvard	M.I.T.	Total
Cambridge (U.S.) (Arrow, Samuelson, Solow)	(19)	(77)	(74)	(133)	(35)	(338)
Favorable	0.0	5.2	9.5	21.1	2.9	11.8
Unfavorable	0.0	11.7	1.4	0.8	5.7	3.8
Cambridge (England) (Kaldor, Robinson)	(10)	(48)	(4)	(46)	(3)	(111)
Favorable	0.0	0.0	0.0	2.2	0.0	0.9
Unfavorable	0.0	0.0	0.0	8.7	0.0	3.6
Monopolistic competition (Bain, Chamberlin, Modigliani, Steiner)	(8)	(38)	(35)	(151)	(4)	(236)
Favorable	12.5	7.9	2.9	13.2	0.0	10.6
Unfavorable	0.0	21.1	2.9	21.9	0.0	17.4
Input-output (Dorfman, Leontief, Solow)	(50)	(13)	(17)	(90)	(4)	(174)
Favorable	10.0	15.4	11.8	21.1	0.0	16.1
Unfavorable	2.0	0.0	0.0	6.7	0.0	4.0
Chicago-Columbia (Becker, Friedman, Stigler)	(2)	(48)	(85)	(49)	(19)	(203)
Favorable	0.0	12.5	4.7	6.1	10.5	7.4
Unfavorable	0.0	2.1	15.3	12.2	36.8	13.3

NOTE.—Number of citations in parentheses.

In addition, a few clear antagonisms are revealed.[31]

We may combine citations of doctorates to the various groups we have distinguished in price theory, and this is done in table 19. It becomes apparent that the input-output area received chief attention from Berkeley and Harvard, and little elsewhere; and the exclusive appeal of monopolistic competition to Harvard is even more striking. The polarity of citations for Harvard and M.I.T. on the one hand and Chicago on the other is conspicuous, with Columbia in an ambiguous position between them.

The monetary authorities seldom fall into well-defined groups. We can distinguish only a few allies:

1. Metzler-Haberler:

$$_1C_2 = 13\ (F = 0;\quad U = 7.7),$$

$$_2C_1 = 10\ (F = 40;\ U = 0).$$

---

[31] Bain-Stigler: $_1C_2 = 8\ (F = 0;\ U = 50)$,
Chamberlin-Kaldor: $_1C_2 = 10\ (F = 10;\ U = 80)$,
Samuelson-Scitovsky: $_1C_2 = 27\ (F = 7.4;\ U = 59.3)$,
Hirshleifer-Friedman $_1C_2 = 18\ (F = 0;\ U = 50)$.

TABLE 20

CITATIONS OF KEYNES BY AUTHORITIES

Citer	N	Favorable (%)	Unfavorable (%)
Lerner	33	6.1	0
Hicks	28	17.9	0
Tobin	20	10.0	0
Harrod	14	21.4	0
Robertson	16	12.5	6.2
Hansen	44	9.1	6.8
Metzler	18	0	0
Smith, W.	12	0	0
Bronfenbrenner	18	5.6	11.1
Friedman	26	11.5	15.4
Johnson	51	3.9	19.6

2. Hicks-Johnson:

$$_1C_2 = 3\ (F = 0;\quad U = 0),$$
$$_2C_1 = 9\ (F = 22.2;\ U = 0).$$

3. Hansen-Samuelson:

$$_1C_2 = 3\ (F = 33.3;\ U = 0),$$
$$_2C_1 = 8\ (F = 37.5;\ U = 0).$$

If we do not require reciprocal quotation,[32] we may tally the allies and opponents of Keynes as shown in table 20. The list illustrates a limitation of citations. The great dispute between Keynes and Robertson was conducted primarily outside the journals.

A comparable tabulation for Friedman (table 21), the leading monetarist, forms an interesting companion. The Keynes and Friedman lists clearly classify Lerner and Tobin, and probably Klein (no citations of Keynes) and Harrod (no citations of Friedman) as Keynesians and nonmonetarists. The monetarists, non-Keynesians are Bronfenbrenner, Brunner, Friedman, Johnson, and Selden. We combine the citations of the doctorates to these two groups in table 22.

The Chicagoans pay great, and only moderately unfavorable, attention to the Keynesians, but all other schools are predominantly favorable. The monetarists are accorded somewhat less attention, but only two schools (Columbia and Harvard) are on balance modestly unfavorable to these authorities. (It is apparent from our tabulations that the English economists in our sample were strongly favorable to Keynes and largely ignored the monetarists.)

Comparison of Citations by Authorities and Doctorates

The citations of the 30-odd leading authorities on price theory and monetary and fiscal theory are available (in reference tables 6 and 7).

---

[32] In the preceding pairs we require at least three citations in each direction.

TABLE 21

CITATIONS OF FRIEDMAN BY AUTHORITIES

Citer	N	Favorable (%)	Unfavorable (%)
Bronfenbrenner......	20	15	0
Selden.............	45	17.8	4.4
Johnson	23	43.5	13.0
Duesenberry	9	33.3	22.2
Brunner	98	11.2	25.5
Tobin	54	14.8	35.2
Lerner	8	12.5	37.5
Klein	16	0	56.2

TABLE 22

CITATIONS BY DOCTORATES TO GROUPS OF ALLIED ECONOMISTS IN MONETARY AND FISCAL ECONOMICS: FAVORABLE AND UNFAVORABLE (PERCENTAGE DISTRIBUTION)

	DOCTORATE					
CITED GROUP	Berkeley	Chicago	Columbia	Harvard	Wisconsin	Total
Keynesians (Keynes, Lerner, Tobin, Klein)...	(66)	(127)	(53)	(93)	(5)	(344)
Favorable...........	22.7	12.6	20.8	12.9	20.0	16.0
Unfavorable.........	7.6	19.7	1.9	8.6	0.0	11.3
Monetarists (Friedman, Bronfenbrenner, Johnson and Selden)....	(15)	(54)	(99)	(37)	(12)	(215)
Favorable...........	23.1	24.1	15.2	16.2	16.7	18.4
Unfavorable.........	7.7	7.4	18.2	21.6	16.7	15.3

NOTE.—Number of citations in parentheses.

These tables, which necessarily exclude economists not writing after 1950[33] and also self-citations, are the bases for the present section.

The citation practices of the authorities display the same heavy influence of propinquity that we find among the doctorates. If we compare the citation of American and English economists by American and English authorities in price theory we find:

	CITATION BY	
CITATION OF	American (%)	English (%)
American.........	71.5	48.9
English...........	28.5	51.1

---

[33] In price theory this includes Marshall, Keynes, Walras, Knight, Pareto, and Schumpeter. In monetary and fiscal economics the exclusions are Clark, Fisher, Hamilton, Keynes, Lange, and Marshall.

An equal measure of chauvinism is found also in the monetary citations.[34] The only question, which we are not presently equipped to answer, is: Which country is chauvinistic? It is a plausible conjecture that the larger the relative size of a group, the more relatively (to size) that group cites itself, but we do not test that hypothesis.

There is, of course, a family resemblance between the citations of the authorities and the doctorates. The simple rank correlation coefficients are .346 for price theory and .625 for monetary and fiscal theory (based on reference tables 6 and 7).[35] The living authorities are considerably older: in 1950 they were about 45 years old on average, the doctorates' average age was 30 years. We find that with the single exception of Irving Fisher the authorities cite the older economists relatively more frequently than the doctorates do. We were inclined to explain some differences between the two rankings by the fact that one-third of the authorities were non-American whereas relatively few doctorates are non-American, but this proved not to be influential.[36]

V. Conclusion

Our foregoing analysis takes the citation as an acceptable statistical measure of influence or attitude, and we believe that this measure deserves deeper and more critical examination than it has so far received. But given its at least approximate validity in this role, citation analysis leads to the conclusion that there are genuine differences among the universities in the attention and respect paid to various scholars. Local faculty receive a perhaps inevitable preference in the students' work,[37] and the ideological preferences of the professors are communicated in some degree to the students.

The main fact, however, is that these influences are not of great strength.

---

[34] The corresponding tabulation is:

CITATION OF	CITATION BY American (%)	CITATION BY English (%)
American	52.4	26.2
English	47.6	73.8

[35] The rank correlation for price theory is slightly lowered by the omission of citations by eight authorities also in the doctorates group; the adjustment would be negligible for monetary theory.

[36] The price and monetary and fiscal correlations for American authorities versus doctorates are .462 and .750, respectively.

[37] We are reminded of the English mathematician G. H. Hardy, who said that to justify his existence a man had to believe in the importance of his work, and of his own capacities in this work (1957, p. 66). These beliefs are calculated to reach the student.

Especially in value theory the differences among schools are small compared with the authorities they share in common. We conjecture that larger differences would be found in the teaching of the other social sciences but presumably even smaller differences in fields such as mathematics. We know of no other studies with which such comparisons can be made, but suggest that this is a method of establishing the uniqueness and strength of a paradigm in a science.

References

American Economic Association. *Index of Economic Articles.* Vols. 1 (1886–1924)–10 (1968). Homewood, Ill.: Irwin, 1961–71.

American Economic Association and Royal Economic Society. *Surveys of Economic Theory.* London: Macmillan; New York: St. Martin's, 1965–66.

Cole, J. R., and Cole, Stephen. *Social Stratification in Science.* Chicago: Univ. Chicago Press, 1973.

Eckstein, O., and Wilson, T. A. "The Determination of Money Wages in American Industry." *Q.J.E.* 76 (August 1962): 379–414.

Ellis, H. S., ed. *A Survey of Contemporary Economics.* Vol. 1. Homewood, Ill.: Irwin, 1948.

Haley, B. F., ed. *A Survey of Contemporary Economics.* Vol. 2. Homewood, Ill.: Irwin, 1952.

Hardy, G. H. *A Mathematician's Apology.* New York: Cambridge Univ. Press, 1957.

Hicks, J. R. "A Rehabilitation of 'Classical' Economics?" *Econ. J.* 67 (1957): 279–89.

Merton, R. *The Sociology of Science.* Chicago: Univ. Chicago Press, 1973.

Tsiang, S. C. "Liquidity Preference and Loanable Funds Theories, Multiplier and Velocity Analyses: A Synthesis." *A.E.R.* 46 (1956): 539–64.

The Demand for Index Bonds

Stanley Fischer

Massachusetts Institute of Technology

The demand for index bonds by households is studied in an intertemporal optimization model in which households can choose a portfolio consisting of equity, nominal bonds, and index bonds. The rate of inflation is uncertain. In a setting where all income is from capital gains, it is shown that index bonds command a premium over nominal bonds only if equity is not a hedge against inflation; it is also shown that, if there are homogeneous expectations and no outside bonds, all lending and borrowing will be through index bonds with no nominal bonds existing. These conclusions are modified when individuals have income from human capital. The very stringent conditions under which a single index bond provides the household with the same utility as a full set of futures markets are examined in a model with many consumption goods.

I. Introduction

The linking of contracts to the price level appears to be a simple and desirable device to reduce the costs of unanticipated inflation. Indexed bonds have been introduced in a number of Latin American and European countries (including, recently and on a limited scale, the United Kingdom) but not, with one exception, in the United States, despite the recommendations of a lengthy list of distinguished economists.[1] The one

---

I am grateful to Robert C. Merton for lengthy discussions and comments; to Zvi Body, Franklin Fisher, Franco Modigliani, and Paul Samuelson of M.I.T.; to Milton Friedman and other members of the University of Chicago Money Workshop; and to Robert Jones for comments and suggestions. Research support from the National Science Foundation is acknowledged with thanks.

[1] A partial list of these economists and the arguments for and against indexed bonds is contained in Ahtiala (1967); see also the highly readable book by Collier (1969). Indexed bonds have been used extensively in Finland and Israel. An account of recent Brazilian experience is contained in Fishlow (1974). Tobin (1971) advances several arguments for federal government issuance of an index bond. The desirability of indexing has also been discussed recently in the daily press, following a suggestion by Friedman (*New York Times*, April 3, 1974).

[*Journal of Political Economy*, 1975, vol. 83, no. 3]

© 1975 by The University of Chicago. All rights reserved.

exception is an issue of index bonds by the Rand Kardex Company in 1925, when, by no coincidence, Irving Fisher was associated with the company.[2] Escalator clauses in wage contracts are used in many countries.

The nonissuance of index bonds in the United States by either the federal government or corporations is something of a mystery. In the case of corporations, the tax treatment of inflation-linked interest payments essentially as dividends may be a major explanatory factor, as may the stability of the inflation rate through the fifties and sixties. There seem to be no mitigating circumstances for the federal government, however; issuance of an indexed bond would, it has been argued, provide portfolio holders with an asset which guards against a risk which cannot now be fully hedged. The existence of such an asset would be particularly desirable on distributional grounds, since it would enable consumers to provide themselves with an assured standard of living for their retirements. Further, since there is an evident demand for the hedge provided by such bonds, they could be sold at a real interest rate below that on nominal bonds, so that no real costs to the Treasury would be involved in the substitution of some indexed bonds for outstanding nonindexed bonds.[3]

The purpose of this paper is to investigate the theory of the demand for index bonds by households. The basic model, outlined in Section II, is of an infinitely lived household receiving income only as a return on asset holdings and able to hold a nominal bond (the nominal return on which is certain), an index bond (which has a certain real return), and equity (the real and nominal returns on which may both be uncertain). There is a single consumption good purchasable at a price P which changes stochastically through time. Demand functions for the three assets are derived, and equilibrium relative yields on the assets and the determinants of the outstanding volume of indexed and nominal bonds, respectively, are examined.

Section III contains a generalization of the model to the situation where there are many goods, as many indexed bonds, each indexed on one of the goods, and many equity assets. The usual mutual fund theorem for equity assets holds. The many indexed bonds can be replaced by a single real asset only if all households have the same utility functions with constant relative risk aversion; if all households have the same expenditure patterns on goods and differing degrees of risk aversion, the many indexed bonds can be replaced by two real assets, one of them a bond indexed on the "ideal" price index.

---

[2] The terms of the bond are described in Pigou (1929, p. 262). The bond was a failure; Fisher attributed the failure to the unfamiliarity of the market with the new asset (see Fisher 1934). However, the indexing of Fisher's bond was peculiar in that a 10 percent change in the price level was required before the indexing provision became effective.

[3] These are essentially the arguments of Tobin (1971), who also suggests that the ability to conduct open-market operations in indexed bonds would increase the Federal Reserve's control over real interest rates and thus potentially improve its stabilization policies.

Section IV returns to the one-good case and elucidates the effects of the existence of wage income on the demand for index bonds and the premium for such bonds in market equilibrium.

Conclusions and unfinished business are summarized in Section V. Appendix 1 contains an informal review of the continuous-time dynamic processes and dynamic-programming techniques used in the paper.[4] A table of notation (Appendix 2) is also provided.

II. The Single-Good, Three-Asset Case

This section contains the basic model of the paper. The model is one of an infinitely lived household receiving all its income as a return on tradable assets and deriving utility from consumption. The household faces a stochastic rate of inflation and can hold a real bond, equity, and a nominal bond in its portfolio. Returns are expressed in nominal terms in this section; this is done partly because generalization to the multigood case is then straightforward and also because, since goods are not generally used as numeraire in practice, empirical work on stock market returns has tended to use nominal rather than real returns.[5]

The major results of the section are contained in parts E and F: asset demands are derived and the properties of the asset demand functions are studied in part E, and market equilibrium and rates of return are discussed in part F.

A. Price Dynamics

The rate of inflation is stochastic, and the behavior of the price level is describable by the stochastic process[6]

$$\frac{dP}{P} = \Pi dt + s dz = \Pi dt + s z(t)\sqrt{dt}. \tag{1}$$

Section 1 of Appendix 1 contains a brief discussion of the meaning of equations such as (1), which states that, over a short time interval, the proportionate change in the price level is normal with mean Πdt and variance $s^2 dt$. The stochastic component of (1) is temporally independent no matter how short the time period considered. It is shown in section 6 of Appendix 1 that (1) implies $P(t)$ is log normally distributed with

---

[4] The stochastic calculus has been used most extensively in the economics literature by Merton (1969, 1971, 1973).

[5] For an exception, see Fama and MacBeth (1974).

[6] It is assumed that the household knows the correct distribution of the price level and returns on assets.

$$E_0\left[\log \frac{P(t)}{P(0)}\right] = \left(\Pi - \frac{s^2}{2}\right)t, \tag{2}$$

$$E_0\left\{\left[\log \frac{P(t)}{P(0)} - \left(E_0 \log \frac{P(t)}{P(0)}\right)\right]^2\right\} = s^2 t, \tag{3}$$

where E_0 is the expectation conditional on $P(0)$. The assumption that Π and s in (1) are constant through time is, of course, very strong, and its relaxation will be discussed briefly in Section V.

B. *Asset Returns*

The household can hold three assets in its portfolio—a real bond, a risky asset (equity), and a nominal bond—and can adjust its portfolio instantaneously and costlessly. There are no nonnegativity constraints on the asset holdings. The real bond pays a real return of r_1 and a nominal return of r_1 plus the realized rate of inflation. Then

$$\frac{dQ_1}{Q_1} = r_1 dt + \frac{dP}{P} = (r_1 + \Pi)dt + sdz \equiv R_1 dt + s_1 dz_1 \tag{4}$$

is the equation describing the nominal return on the index bond.[7]

The nominal return on equity is[8]

$$\frac{dQ_2}{Q_2} = R_2 dt + s_2 dz_2, \tag{5}$$

where R_2 is the expected nominal return on equity per unit time and s_2^2 is the variance of the nominal return per unit time. It will later turn out that the covariance of the nominal return on equity with that on real bonds—and so with the rate of inflation—is a key variable in the demand functions for assets. By Itô's lemma (Appendix 1, sec. 4) that covariance is $\rho s_1 s_2 dt$, where ρ is the instantaneous coefficient of correlation between the Wiener processes, dz_1 and dz_2 ($|\rho| < 1$).

---

[7] The return on the real bond may be regarded as accruing either as an increase in the price of a single unit of the bond or as an increase in the number of units of fixed nominal value held by the household. The return on equity is most usefully thought of as an increase in the price of the unit of equity (see Merton 1973, pp. 870–71).

[8] The assumption of only one equity asset is innocuous, since a generalized separation theorem holds in the presence of many risky assets. Namely, given a nominal bond, real bond, and many equity assets, each with returns which are log normally distributed, any household will be indifferent between being given the opportunity of holding all the assets and the opportunity of holding the nominal bond, real bond, and a specified mutual fund of equity assets. Alternatively, the equity assets will be held in the same proportions (to each other) in all portfolios, given agreement on the distribution of returns (see Sec. III below).

The *real* return on equity is

$$\frac{d(Q_2/P)}{Q_2/P} = (R_2 - \Pi - \rho s_1 s_2 + s_1^2)dt + s_2 dz_2 - s_1 dz_1 \qquad (6)$$
$$= r_2 dt + s_2 dz_2 - s_1 dz_1.$$

Once again Itô's lemma is used in deriving (6) in a manner similar to that shown in section 5 of Appendix 1. Note particularly that the expected real return on equity, r_2, is not the expected nominal return minus the expected rate of inflation. This latter relationship would hold if there were no uncertainty about the rates of inflation and return on equity or, alternatively, if $\rho s_2 = s_1$, that is, if there were zero covariance between the real returns on equity and nominal bonds (see eq. [7] below). It does not hold in general, however.

The covariance of the real return on equity with the rate of inflation is $\rho s_1 s_2 - s_1^2$, and the coefficient of correlation between the real return on equity and the real return on nominal bonds

$$\lambda = -\frac{\rho s_2 - s_1}{\sqrt{s_1^2 - 2\rho s_1 s_2 + s_2^2}}. \qquad (7)$$

The deterministic nominal return on nominal bonds is

$$\frac{dQ_3}{Q_3} = R_3 dt. \qquad (8)$$

The real return on the nominal bond is shown in section 5 of Appendix 1 to be

$$\frac{d(Q_3/P)}{Q_3/P} = (R_3 - \Pi + s_1^2)dt - s_1 dz_1 = r_3 dt - s_1 dz_1. \qquad (9)$$

Again, the expected real return on the nominal bond, r_3, is not the nominal rate minus the expected rate of inflation, though that would be the case if the rate of inflation were not stochastic.[9]

[9] This result has to do with the fact that, in general,

$$E\left[\frac{1}{x}\right] \neq \frac{1}{E[x]}.$$

For instance, consider a situation in which $P_t = 1$ and $P_{t+1} = 2$ with probability .5 and $P_{t+1} = 0.5$ with probability .5. Then $E[P_{t+1}] = 1.25$ and the expected rate of inflation

$$E\left[\frac{P_{t+1}}{P_t} - 1\right] = 0.25.$$

Obviously, however,

$$E\left[\frac{1}{P_{t+1}}\right] = 1.25,$$

so that the expected rate of deflation

$$E\left[\frac{P_t}{P_{t+1}} - 1\right] = 0.25.$$

Thus, the expected return from holding money, paying no interest, is positive even though the expected rate of inflation is positive.

C. *Budget Constraints*

Let w_1, w_2, and w_3 be the proportions of the portfolio held in real bonds, equity, and nominal bonds, respectively. Then the stock budget constraint is obviously

$$1 = w_1 + w_2 + w_3. \tag{10}$$

The flow budget constraint, giving the change in nominal wealth (W), is

$$dW = \sum_{i=1}^{3} w_i R_i W dt - PCdt + \sum_{i=1}^{2} w_i s_i W dz_i, \tag{11}$$

where C is the rate of consumption. Uncertainty about the change in nominal wealth arises from holdings of real bonds and equity and is described by the last two terms on the right-hand side of (11). Substituting for w_3 from (10) into (11) gives

$$dW = \sum_{i=1}^{2} w_i(R_i - R_3)W dt + (R_3 W - PC)dt + \sum_{i=1}^{2} w_i s_i W dz_i. \tag{12}$$

D. *The Household's Choice Problem*

The household is to find

$$\max_{\{C, w_i\}} E_0 \int_0^{\infty} U[C(t), t]dt \tag{13}$$

subject to (12) and $W(0) = W_0$ and with $U(\;)$ strictly concave in C. The solution to the choice problem is found by solving for the derived utility function

$$J(W, P, t) = \max_{\{C, w_i\}} E_t \int_t^{\infty} U[C(\tau), \tau)d\tau. \tag{14}$$

The general method of solving for $J(\;)$ is stated in sections 7 and 8 of Appendix 1.

Using this method, we obtain the first-order conditions

$$0 = U_c(C, t) - PJ_W, \tag{15}$$

$$0 = J_W(R_1 - R_3) + J_{WW}W(w_1 s_1^2 + w_2 \rho s_1 s_2) + J_{WP}Ps_1^2, \tag{16}$$

$$0 = J_W(R_2 - R_3) + J_{WW}W(w_1 \rho s_1 s_2 + w_2 s_2^2) + J_{WP}P\rho s_1 s_2. \tag{17}$$

Equation (15) is simply the condition that the marginal utility of consumption per dollar be equated to the marginal utility, in the derived utility function, of nominal wealth. Equations (16) and (17) are similar to the usual portfolio equations except for the terms in J_{WP}.

E. Asset Demands

Solving (16) and (17) for w_1 and w_2 gives

$$w_1 = -\frac{J_W}{J_{WW}W}\left[\frac{R_1 - R_3}{s_1^2(1-\rho^2)} - \frac{\rho(R_2 - R_3)}{s_1 s_2(1-\rho^2)}\right] - \frac{J_{WP}P}{J_{WW}W}, \quad (18)$$

$$w_2 = -\frac{J_W}{J_{WW}W}\left[\frac{R_2 - R_3}{s_2^2(1-\rho^2)} - \frac{\rho(R_1 - R_3)}{s_1 s_2(1-\rho^2)}\right]. \quad (19)$$

In interpreting these equations, we consider first the term involving $J_{WP}P$ in (18). Differentiating the first-order condition (15) with respect to P and then with respect to W, we obtain

$$U_{cc}\frac{\partial C}{\partial P} = J_W + PJ_{WP},$$

$$U_{cc}\frac{\partial C}{\partial W} = PJ_{WW}.$$

Then, since consumption is a function of real wealth, note that

$$\frac{\partial C}{\partial P} = -\frac{W}{P}\frac{\partial C}{\partial W}.$$

Hence,

$$\frac{J_{WP}P}{J_{WW}W} = -\frac{J_W}{J_{WW}W} - 1. \quad (20)$$

Thus, we obtain

$$w_1 = 1 - \frac{J_W}{J_{WW}W}\left[\frac{R_1 - R_3}{s_1^2(1-\rho^2)} - \frac{\rho(R_2 - R_3)}{s_1 s_2(1-\rho^2)} - 1\right]. \quad (18')$$

Consider now the demand function for index bonds (eq. [18′]).[10] The coefficient of the term in brackets $(-J_W/J_{WW}W)$ is the inverse of the degree of relative risk aversion of the household and will henceforth be denoted $A(W, P)$. Assume for the moment that $\rho = 0$. Then the demand for index bonds depends in an intuitively plausible way on the degree of relative risk aversion, the difference between expected nominal returns on the two types of bonds $(R_1 - R_3)$, and the variance of the price level (s_1^2).

---

[10] It is possible to move from (18′) and (19) to demand functions for assets expressed in terms of means and variances of real returns. Such demand functions are presented in Section IV below. The reasons for presenting the analysis in nominal terms are discussed above.

It is useful at this point to compute the demand function for nominal bonds using (18′), (19), and (10):

$$w_3 = -A(W, P)\left[\frac{(R_1 - R_3)(s_2 - \rho s_1)}{s_1^2 s_2(1 - \rho^2)} + \frac{(R_2 - R_3)(s_1 - \rho s_2)}{s_1 s_2^2(1 - \rho^2)} - 1\right]. \tag{21}$$

The complete properties of the demand functions for the three assets are specified in tables 1 and 2. The demand for each asset is positively related to its own expected nominal rate of return; equity and the index bond are substitutes in that the demand for each is negatively related to the expected nominal return on the other (for $\rho > 0$). The other derivatives of the demand functions are, in general, of ambiguous sign. If the covariance of real returns on equity and nominal bonds is positive (i.e., $\rho s_2 - s_1 > 0$), then equity and nominal bonds are substitutes with respect to expected nominal rates of return.

The sources of the indeterminacies in the asset demand functions can be isolated with the help of table 2. If $\rho = 0$, the assets are substitutes with respect to expected nominal rates of return (except for equity and the index bond, which are borderline between being complements and substitutes for each other). Thus, the source of the ambiguity in the effects of changes in the nominal interest rate on the demands for index bonds and equity is revealed to be the correlation between the nominal returns on those assets, which permits equity to serve as a hedge against inflation (in terms of expected nominal returns). Similarly, for $s_1 = \rho s_2$, the assets are substitutes with respect to expected nominal returns (except that now equity and nominal bonds are borderline).

The ordering of expected returns assumed in case C of table 2 ($R_2 > R_3 > R_1$) removes the ambiguities in the effects of s_1, s_2, and ρ on the demands for the index bond and equity. However, note that although this ordering seems natural, it is not necessarily implied by the household's owning positive amounts of all three assets. In case D of table 2, for both $\rho = 0$ and $R_2 > R_3 > R_1$, all but one of the signs are determinate and readily explicable.

F. Rates of Return on Indexed Bonds and Market Equilibrium

We now discuss the rates of return, real and nominal, required on index bonds if the household is to hold any of the asset, and also the determinants of the gross quantity of outstanding indexed bonds.[11] In discussing rates of return required for a particular household to hold the

---

[11] I am indebted to Franco Modigliani for pointing out the consequences for market equilibrium of the forms of the demand functions for assets obtained above; these points are clarified in Modigliani (1974).

TABLE 1

PROPERTIES OF THE ASSET DEMAND FUNCTIONS*

ASSET	PARAMETER					
	R_1	R_2	R_3	s_1	s_2	ρ
w_1	$\frac{1}{\xi s_1^2}$	$-\frac{\rho}{\xi s_1 s_2}$	$\frac{\rho s_1 - s_2}{\xi s_1 s_2^2}$	$\frac{\rho s_1 b - 2s_2 a}{\xi s_1^3 s_2}$	$\frac{\rho b}{\xi s_1 s_2^2}$	$\frac{2\rho s_2 a - s_1(1+\rho^2)b}{\xi(1-\rho^2)s_1^2 s_2}$
Index bond....	+	−	...	...	...	...
w_2	$-\frac{\rho}{\xi s_1 s_2}$	$\frac{1}{\xi s_2^2}$	$\frac{\rho s_2 - s_1}{\xi s_1 s_2^2}$	$\frac{\rho a}{\xi s_1^2 s_2}$	$\frac{\rho s_2 a - 2s_1 b}{\xi s_1 s_2^3}$	$\frac{2\rho s_1 b - s_2(1+\rho^2)a}{\xi(1-\rho^2)s_1 s_2^2}$
Equity........	−	+	...	...	...	...
w_3	$\frac{\rho s_1 - s_2}{\xi s_1^2 s_2}$	$\frac{\rho s_2 - s_1}{\xi s_1 s_2^2}$	$\frac{s_2^2 - 2\rho s_1 s_2 + s_1^2}{\xi s_1^2 s_2^2}$	$\frac{2s_2 a - \rho s_1(a+b)}{\xi s_1^3 a_2}$	$\frac{2s_1 b - \rho s_2(a+b)}{\xi s_1 s_2^3}$	$\frac{s_1 s_2(1+\rho^2)(a+b) - 2\rho(s_1^2 b + s_2^2 a)}{\xi(1-\rho^2)s_1^2 s_2^2}$.
Nominal bond .	...	...	+	...	...	....

NOTE.—(i) $\xi = \frac{1}{A(W, P)}(1 - \rho^2)$; (ii) ρ is assumed positive; (iii) $a = R_1 - R_3$; (iv) $b = R_2 - R_3$

* Entries are the partial derivatives of the demand for the asset (row) with respect to a parameter (column). Plus and minus signs indicate sign of the derivative. Derivative is of ambiguous sign in other cases.

TABLE 2

PROPERTIES OF THE ASSET DEMAND FUNCTIONS FOR SPECIAL CASES*

ASSET	PARAMETER					
	R_1	R_2	R_3	s_1	s_2	ρ
Case A ($\rho = 0$):						
w_1	+	0	−	?	0	?
w_2	0	+	−	0	?	?
w_3	−	−	+	?	?	?
Case B ($s_1 = \rho s_2$):						
w_1	+	−	−	?	?	?
w_2	−	+	0	?	?	?
w_3	−	0	+	?	?	?
Case C ($R_2 > R_3 > R_1$):						
w_1	+	−	?	+	+	−
w_2	−	+	?	−	−	+
w_3	?	?	+	?	?	?
Case D ($\rho = 0$ and $R_2 > R_3 > R_1$):						
w_1	+	0	−	+	0	−
w_2	0	+	−	0	−	+
w_3	−	−	+	−	+	?

* Entries indicate the signs of the partial derivatives of the asset demands with respect to the parameters; "?" means that the sign is ambiguous.

assets, we are also implicitly discussing equilibrium relative rates of return in a market with portfolio holders of identical tastes and fixed supplies of assets.

The required yield differentials are examined at the point $w_1 = 0$, that is, where the household has no index bonds in its portfolio. Then, using (18′) and (19) yields[12]

$$R_1 - R_3 = s_1^2 + \frac{1}{A(W, P)} (\rho s_1 s_2 w_2 - s_1^2), \tag{22}$$

and in terms of real returns (using [4] and [9]), we have

$$r_1 - r_3 = \frac{1}{A(W, P)} [(\rho s_1 s_2 - s_1^2) w_2 - s_1^2 (1 - w_2)]. \tag{23}$$

Suppose now that $w_2 = 1$, that is, the net quantities of both types of bond are zero. Then, from (23), the difference between the required real yields depends only on the covariance of the real returns on equity with the rate of inflation ($\rho s_2 - s_1$). If there is positive covariance between equity returns and inflation, then $r_1 > r_3$, that is, index bonds have to pay a higher return than the expected real yield on nominal bonds.[13] Put differently, if equity is a hedge against inflation, then index bonds do not command a premium over nominal bonds; if equity is not a hedge against inflation, index bonds will be sold at a premium over nominal bonds (where, by definition, there is a premium only if $r_1 < r_3$). The effect of changes in w_2 on the premium depends only on the sign of ρ, the correlation coefficient between nominal returns on equity and the price level. For $\rho > 0$, reductions in w_2, equivalent to increases in the outstanding stock of nominal bonds issued, say, by the government, increase the premium of indexed over nominal bonds.

---

[12] The forms of (22) and (23) show clearly that, even given the existence of a real bond, the real yield on nominal bonds is not the nominal yield minus the expected rate of inflation. From (23), if there is no risk aversion, $r_1 = r_3$, i.e., expected real yields on the two types of bonds are equal. However, this does not imply $R_1 = R_3$, rather $R_1 = R_3 + s_1^2$. With risk aversion, even the latter equality does not hold. Thus, the well-known Fisher equation for real rates becomes difficult to interpret. For a fuller discussion, considering also the situation in which there is no indexed bond, see Fischer (1974). See also Tobin (1965) and Sarnat (1973).

It is noteworthy, however, from (22) that even for $w_2 = 1$ and for zero covariance of real returns on equity with inflation, $R_1 - R_3 = s_1^2$, that is, nominal bonds command a premium over real bonds in terms of expected nominal returns. The reason is that the expected real return on nominal bonds is not the nominal interest rate minus the expected rate of inflation (see eq. [9]).

Summarizing the last two paragraphs: if $w_2 = 1$ and $w_1 = 0 = w_3$, then there is a premium for index bonds in terms of expected real returns (i.e., $r_1 < r_3$) only if equity is not a hedge against inflation in real terms (i.e., only if $\lambda > 0$). However, even if there is a premium for index bonds in terms of real returns, there may not be a premium in terms of nominal returns, due to the presence of the s_1^2 term of equation (9) in translating between expected real and nominal returns on nominal bonds. The extent of the premium is also affected by the outstanding stock of nominal bonds.

Although the premium for indexed bonds, positive or negative, is of interest—particularly from the viewpoint of government introduction of such bonds—information on the premium does not directly bear on the quantities of the two types of bonds likely to exist in private markets. Assume there are no outside assets in existence so that in aggregate $w_1 = 0$, $w_2 = 1$, and $w_3 = 0$. It is helpful at this point to refer back to equation (21), the demand function for nominal bonds.

It is clear from (21) that, if expectations are identical across individuals, then each individual's demand for nominal bonds is simply some constant—the same across individuals—multiplied by the negative of $[A(W, P)]$. Thus, if everyone is risk averse, every individual's demand for nominal bonds is of the same sign. Since the net demand has to be zero across the market, each individual's demand must be zero. Accordingly, equilibrium rates of return adjust so that no individual is either borrowing or lending through nominal bonds, even if individuals' risk aversions differ. Notice that this in no way contradicts the fact that nominal bonds might command a premium over index bonds—they are simply priced at precisely that rate which ensures that none of them exist.

---

[13] Modigliani (1974) points out that similar conclusions can be reached by a direct application of the capital asset pricing model.

Examination of (18′) and (19) reveals that the same argument does not apply for indexed bonds. The outstanding quantity of indexed bonds will depend on the dispersion of attitudes to risk: more risk averse individuals hold positive amounts of indexed bonds, and less risk averse individuals borrow, issuing indexed bonds to buy equity.

However, if expectations are not identical, there may be positive quantities of nominal bonds outstanding: those who expect relatively low rates of inflation holding nominal bonds and those who expect relatively high rates issuing them. Nonetheless, the result that nominal bonds would not exist in this model with homogeneous expectations is striking and will be discussed again in Section V.

III. The Multigoods, Multiindexed Bonds Case

The analysis of the demand for indexed bonds above assumes the existence of only one good or, equivalently, of fixed relative prices of goods. In the presence of many different goods whose relative prices change, there is an obvious difficulty in choosing the index on the basis of which real contracts are made. Rather than directly confront the problem of the conditions under which an appropriate index for an index bond can be defined, we proceed indirectly by first assuming the existence of as many indexed bonds as goods, each bond indexed to a single good—in effect, future markets in each good. We then consider the conditions—which turn out to be surprisingly stringent—under which a mutual fund of the indexed bonds can be created such that all households are indifferent between holding the full set of assets or a set of assets in which there is only a single indexed bond.[14]

Let the price dynamics for each good be

$$\frac{dP_i}{P_i} = \Pi_i dt + s_i dz_i, \qquad i = 1, \ldots, m, \tag{24}$$

and let the first m assets be indexed bonds, each paying nominal returns of

$$\frac{dQ_i}{Q_i} = r_i dt + \frac{dP_i}{P_i} = (r_i + \Pi_i)dt + s_i dz_i \equiv R_i dt + s_i dz_i, \qquad i = 1, \ldots, m. \tag{25}$$

---

[14] Portfolio problems with many goods have been studied by Roll (1973), Jones (1974), and Long (1974).

There are $n - m$ equity assets with nominal returns dynamics of

$$\frac{dQ_j}{Q_j} = R_j dt + s_j dz_j, \qquad j = m + 1, \ldots, n, \tag{26}$$

and a nominal bond yielding returns of

$$\frac{dQ_{n+1}}{Q_{n+1}} = R_{n+1} dt \equiv R dt. \tag{27}$$

The covariances of nominal returns on the ith and jth assets are given by $\rho_{ij}s_is_j$; the assumed nonsingular variance-covariance matrix of returns on the first n assets is defined by

$$\Omega = [\rho_{ij}s_is_j], \qquad i = 1, \ldots, n; j = 1, \ldots, n; \rho_{ii} = 1. \tag{28}$$

The household now maximizes

$$E_0 \int_0^\infty U(C_1, \ldots, C_m, t)dt, \tag{29}$$

where $U(\ )$ is strictly concave in the C_i subject to

$$dW = \sum_{i=1}^{n} w_i(R_i - R)Wdt + \left(RW - \sum_{i=1}^{m} P_iC_i\right)dt + \sum_{i=1}^{n} w_is_iWdz_i. \tag{30}$$

The optimization method is as in Section II and involves finding the derived utility function $J(P_1, \ldots, P_m, W, t)$. Define the vectors

$$\begin{aligned} [\hat{R} - R] &= [R_i - R \quad R_j - R]' \\ &\equiv [R_1 - R \ldots R_m - R \quad R_{m+1} - R \ldots R_n - R]', \\ [w] &= [w_i \quad w_j]' \equiv [w_1 \ldots w_m \quad w_{m+1} \ldots w_n]', \\ [J_{WP}P] &= [J_{WP_i}P_i \quad 0]' \equiv [J_{WP_1}P_1 \ldots J_{WP_m}P_m \quad 0 \ldots 0]'. \end{aligned}$$

The first-order conditions yield the asset demand functions

$$\begin{bmatrix} w_i \\ w_j \end{bmatrix} = -\frac{J_W}{J_{WW}W}\Omega^{-1}\begin{bmatrix} R_i - R \\ R_j - R \end{bmatrix} - \frac{1}{J_{WW}W}\begin{bmatrix} J_{WP_i} & P_i \\ 0 & \end{bmatrix}. \tag{31}$$

It is apparent from inspection of (31) that, in the presence of the m real bonds, a single-equity mutual fund can be formed such that, if expectations are identical, all households will be indifferent between portfolios in which they have a choice of holding the original $n + 1$ assets and those in which they hold the m real bonds, the nominal bond, and the equity mutual fund.

Examination of (31) reveals two elements in the demand for indexed bonds: one is a hedging demand, represented by the term in J_{WP_i}; the other is a speculative demand in which expected returns and variances have precisely the same effect as they do in the demand for any other asset.[15]

We next consider the conditions under which a single indexed bond can be created to provide the household with the same pattern of returns as the original in indexed bonds. The question can be posed by asking whether a "real" mutual fund can be formed out of the indexed bonds such that the household is indifferent between the original asset menu and one consisting only of the equity mutual fund, real mutual fund, and nominal bond. It can be shown that such a real mutual fund, with the proportions of the original bonds in the fund independent of wealth, can be formed for the household if and only if its utility function has constant relative risk aversion (CRRA).[16] Thus, in general such a mutual fund cannot be created even for an individual household.

Further, although a mutual fund for an individual household exists in such cases, different mutual funds are required for different households unless all households (already assumed to have CRRA utility functions) have both the same expenditure shares on goods and the same degrees of relative risk aversion. However, if households have the same expenditure shares and different degrees of risk aversion, a third mutual fund can be created such that all households are indifferent between the original asset menu and one consisting of the nominal bond, equity mutual fund, a bond formed of the original indexed bonds in the same proportions as expenditure shares (a hedging fund), and a speculative real fund formed in the obvious way from (31) to satisfy the speculative demand defined above.[17] Once again, though, in general a single or even two real bonds, in the form of mutual funds, do not provide the household with the same opportunities as the full range of indexed bonds.

Thus, the ability of a single indexed bond to provide households with the same opportunities as a full set of indexed bonds is very restricted. Even under the favorable—and unrealistic—conditions in which all households have identical expenditure shares and constant relative risk aversion, a single indexed bond will not in general provide the same opportunities for all households as a full set of indexed bonds. In this case, however, two indexed bonds, one indexed on the appropriate "ideal" price level[18] and the other created to satisfy the speculative demands for commodities futures, would do so. The extent to which a single indexed bond "nearly" provides the same opportunities for all households as the full set of bonds obviously depends on how close to constant relative risk aversion are utility functions, the similarity of tastes among households, the correlations of price changes, and the risk aversion of households.[19]

---

[15] Incidentally, the results that there would be no nominal bonds in existence in market equilibrium and that indexed bonds do not necessarily command a premium over nominal bonds hold here also.

[16] See Hanoch (1974) for utility functions with constant relative risk aversion in the multigoods case.

[17] I am grateful to Robert Jones for this point.

[18] See Samuelson and Swamy (1974).

IV. Wage Income

In this section we return to the one-commodity–one-equity asset framework to consider the effects of the existence of uncertain wage income on the demands for assets in general and for index bonds in particular. To simplify the analysis, all returns in this section are expressed in real terms.[20] Translation between real and nominal terms is easily accomplished with the aid of equations (4), (6), (7), and (9) of Section II, together with (35′) below.

The behavior of the rate of inflation continues to be described by (1). The real returns on real bonds, equity, and nominal bonds, respectively, are given, using equations (4), (6), and (9), as

$$\frac{d(Q_1/P)}{Q_1/P} = r_1 dt, \tag{32}$$

$$\frac{d(Q_2/P)}{Q_2/P} = r_2 dt + s_2 dz_2 - s_1 dz_1 \equiv r_2 dt + \sigma_2 dx_2, \tag{33}$$

$$\frac{d(Q_3/P)}{Q_3/P} = r_3 dt - s_1 dz_1 \equiv r_3 dt + \sigma_3 dx_3. \tag{34}$$

The coefficient of correlation between the Wiener processes dx_2 and dx_3 is denoted λ, as in (7).

The stochastic behavior of real labor income is also modeled as an Itô process:[21]

$$\frac{dY}{Y} = v dt + b dq. \tag{35}$$

Let η be the coefficient of correlation between real wage income and the rate of inflation. Denote by η_2 and η_3 $(= -\eta)$ the coefficients of correlation between the Wiener process dq and the processes dx_2 and dx_3

---

[19] The hedging demand becomes more important relative to the speculative demand the greater the household's risk aversion.

[20] Thus, the demand functions (36)–(38) below, with J_{YY} set to zero, are the counterparts of (18′), (19), and (21), expressed now in terms of expected real returns and variances.

[21] Merton (1971) solves a problem in which the behavior of wage income is described by a Poisson process. Siegel (1974) discusses the role of human capital in the demand for index bonds.

(eqs. [33] and [34]), respectively; that is, η_2 is the coefficient of correlation between real labor income and real equity return, and η_3 is the coefficient of correlation between real labor income and the real return on nominal bonds.

Letting w_1, w_2, and w_3 once again be the portfolio shares of the index bonds, equity, and nominal bonds, respectively, with V now as real wealth and $A(V, Y)$ the inverse of the degree of relative risk aversion, the demand functions for the three assets are

$$w_2 = A(V, Y)\left[\frac{r_2 - r_1}{\sigma_2^2(1 - \lambda^2)} - \frac{\lambda(r_3 - r_1)}{\sigma_2\sigma_3(1 - \lambda^2)}\right] - \frac{J_{VY}Y}{J_{VV}V}\frac{b}{\sigma_2(1 - \lambda^2)}[\eta_2 - \lambda\eta_3], \tag{36}$$

$$w_3 = A(V, Y)\left[\frac{r_3 - r_1}{\sigma_3^2(1 - \lambda^2)} - \frac{\lambda(r_2 - r_1)}{\sigma_2\sigma_3(1 - \lambda^2)}\right] - \frac{J_{VY}Y}{J_{VV}V}\frac{b}{\sigma_3(1 - \lambda^2)}[\eta_3 - \lambda\eta_2], \tag{37}$$

$$w_1 = 1 - A(V, Y)\left[\frac{(r_2 - r_1)(\sigma_3 - \lambda\sigma_2)}{\sigma_2^2\sigma_3(1 - \lambda^2)} + \frac{(r_3 - r_1)(\sigma_2 - \lambda\sigma_3)}{\sigma_2\sigma_3^2(1 - \lambda^2)}\right] + \frac{J_{VY}Y}{J_{VV}V}\frac{b}{\sigma_2\sigma_3(1 - \lambda^2)}[(\eta_3 - \lambda\eta_2)\sigma_2 + (\eta_2 - \lambda\eta_3)\sigma_3]. \tag{38}$$

At first glance, comparing (36)–(38) with similar equations for the no-wage-income case, one is tempted to conclude that the existence of wage income affects the asset demands only through the terms in J_{VY}, the hedging terms. These hedging terms arise from the nontradability of human capital, which implies that the riskiness of wage income can be diversified away only by holding other assets. However, there is also a wealth effect produced by the existence of wage income which, appropriately capitalized, is treated as part of wealth in the asset demand functions.[22]

To study the hedging effect on the demand for index bonds, consider the interpretation and sign of the term in J_{VY} in (38). From the first-order condition for consumption, $U_C = J_V$, one can show that

$$\frac{J_{VY}Y}{J_{VV}V} = \frac{(Y/C)(\partial C/\partial Y)}{(V/C)(\partial C/\partial V)},$$

---

[22] The nature of these two effects can be seen precisely for the CRRA case in Fischer (1974).

so that the coefficient in the hedging terms is simply the ratio of the income elasticity to the wealth elasticity of consumption demand. It is natural to assume that both $\partial C/\partial Y$ and $\partial C/\partial V$ are positive. Assume also, for this paragraph, that $\lambda = 0$ and, to begin, that $\eta_2 = 0$—that equity returns are uncorrelated with the real wage. Then the sign of the hedging term in (38) is the same as η_3 $(= -\eta)$, the correlation between real labor income and the real return on nominal bonds. For $\eta_3 > 0$, that is, if real labor income is negatively correlated with the rate of inflation, the hedging term is positive. Thus, if the major uncertainty about real wage income arises from fluctuations in the purchasing power of a relatively stable nominal wage income, the existence of wage income produces a positive hedging effect on the demand for index bonds. If the return on equity is positively correlated with real wage income, the hedging effect on the demand for index bonds is increased. Put differently, and somewhat loosely, if changes in real labor income occur largely through changes in labor's share and so are negatively correlated with the returns on equity, the existence of wage income does not necessarily increase the demand for index bonds.

The general principle for the hedging term is that the existence of wage income shifts portfolio demands in the direction of assets the holding of which offsets the riskiness of the real return on human capital.

We turn now to the equilibrium yield, in real terms, on index bonds in the presence of wage income. To do so, we examine the equilibrium yield relationships at the point $w_1 = 0 = w_3$, $w_2 = 1$. Then

$$r_1 - r_3 = -\frac{\sigma_3}{A(V, Y)}\left(\lambda\sigma_2 + \frac{J_{VY}Y}{J_{VV}V} b\eta_3\right). \tag{39}$$

Thus, for $\lambda = 0$, if $\eta_3 > 0$, the existence of wage income means that there will be a premium for real bonds over nominal bonds at the point $w_1 = 0$, $w_2 = 1$. If $\eta_3 > 0$, then there is a positive correlation between the real returns on nominal bonds and the real wage, that is, the real wage is negatively correlated with the rate of inflation; this negative correlation tends to produce a premium for real bonds over nominal bonds. Obviously, however, if $\eta_3 < 0$ and the real wage is positively correlated with inflation, then with $\lambda = 0$, there is a premium for nominal bonds over index bonds. It is perhaps also worth repeating that it is not necessarily true that $r_3 > r_1$ implies $R_3 > R_1$ (see [9]); a premium of index bonds in terms of real returns is not necessarily a premium in nominal terms.

Examination of (37) shows that the striking result of Section II that, with homogeneous expectations and no outside assets, there would be no nominal bonds in existence no longer holds in the presence of human capital if the stochastic properties of wage income differ among individuals.

V. Summary and Conclusions

1. In the simplest model, there is not necessarily a premium for indexed bonds over nominal bonds, in the sense that the real interest rate required to induce a portfolio holder to hold an indexed bond is below the expected real rate on nominal bonds. There will be such a premium if real returns on equity are negatively correlated with the rate of inflation, and there will not be such a premium if real equity returns are positively correlated with the rate of inflation, that is, if equity is a hedge against inflation.[23]

However, it is possible for there not to be a premium for index bonds over nominal bonds in terms of nominal returns even if there is one in terms of real returns. This is because the expected real return on nominal bonds exceeds the nominal rate minus the expected rate of inflation by an amount equal to the variance, per unit time, of the rate of inflation.

In market equilibrium, with zero net supplies of both bonds and homogeneous expectations, all borrowing and lending—arising only from differences in attitudes toward risk—would be through index bonds and none through nominal bonds.

2. Where there are many goods and many indexed bonds, there are two sources of demand for each bond. One is a hedging demand, related to the share of that good in the consumption basket, and the other is a speculative demand, which tends to increase the demand for bonds indexed on the prices of goods expected to rise relatively rapidly, given equal real rates on all bonds. The presence of the speculative demand means that the indexed bonds are not held in the portfolio in the same proportions as their expenditure shares; thus, even for utility functions for which an "ideal" price index can be computed, a single bond indexed on the ideal index is not the ideal indexed bond. However, with identical expenditure shares and constant relative risk aversion for all households, two real assets, one a bond indexed on the ideal index, can provide the same opportunities for each household as a full set of indexed bonds.

3. The effects of wage income—the return on human capital—on the demand for indexed bonds depend on the covariances of wage income with the rate of inflation and equity returns. The premium for index bonds tends to be greater the smaller (algebraically) the covariance of real equity returns with the rate of inflation and the smaller (algebraically) the covariance of real wage income with the rate of inflation. Thus, if human capital is essentially a real asset, so that real wage income has

---

[23] These statements hold exactly only at the point where net outstanding stocks of both real and nominal bonds are zero. The more general conditions are given in Section I, part F. Tobin, in his classic essay (1971), suggests that real equity returns are likely to have zero covariance with the rate of inflation. See also n. 25 below.

zero covariance with the rate of inflation, the existence of wage income does not increase the premium of real bonds over index bonds. If human capital is in some respects a nominal asset, so that real wage income is negatively correlated with the rate of inflation, then wage income tends to increase the premium of index over nominal bonds.[24]

Further, in market equilibrium, with human capital, both types of bonds would in general exist even with homogeneous expectations and no net supplies of the two types of bonds.

4. Analysis of the welfare effects of the government introduction of index bonds would require the specification of a more complete model, with a government budget constraint, than is contained in this paper. Obviously, no household can be made worse off by an expansion of its opportunity set; however, government introduction of an indexed bond would change rates of return on other assets and could make the representative household in market equilibrium worse off.

5. Reasons for the nonexistence of indexed bonds in the U.S. economy do not emerge clearly from the analysis, since in none of the models do nominal bonds drive index bonds out of existence. Other factors, such as tax treatment of indexed payments and the relative stability of the U.S. inflation rate in conjunction with the costs of introducing a new asset and the difficulty of choosing a single appropriate index, would have to provide the explanation for the failure of private borrowers to issue index bonds.

6. Conclusions 1–3 point to the key role of the correlation of equity returns with the rate of inflation in determining the market price at which indexed bonds will be sold. Essentially, indexed bonds will command a real premium if real returns on equity have negative correlation with the rate of inflation. The magnitude of that correlation is, of course, an empirical matter.[25] Nothing in the analysis suggests that the value of nominal bonds would be destroyed if index bonds are introduced at a time when nominal bonds are outstanding.

Factors omitted from the basic model might systematically increase the demand for index bonds relative to the demand functions derived in this paper. Among these are:

a) We have assumed that expectations are correct. The real return on indexed bonds would be known ex ante, while the household has only a (possibly incorrect) probability distribution of returns on nominal bonds.

---

[24] These statements are made for the point at which net quantities of real and nominal bonds are both zero; each statement assumes real equity returns have zero correlation with the rate of inflation.

[25] A preliminary report on research on this question for a number of countries is contained in Cagan (1974), who finds that equity appears to be a long-term, but not short-term, hedge against inflation. Zvi Body, of M.I.T., is currently examining this question with U.S. data.

To the extent that inflation is underpredicted, holders of nominal bonds are made worse off by errors in expectations, while holders of indexed bonds are not.

b) The fact that equity assets are not directly held by the majority of households may indicate that the transactions costs of acquiring information about and purchasing equity are large enough to leave many savers without any potential inflation hedges in their portfolios, aside, perhaps, from real estate. If an indexed bond were made available by the government, say in the form of indexed U.S. Savings Bonds, it might be precisely those savers who would demand it and who would most benefit from it.

c) It was assumed that the stochastic processes generating the rates of inflation and asset returns were stationary. The instantaneous nominal interest rate for all future periods, for instance, was assumed known. Needless to say, this is unrealistic. It is possible that too little uncertainty about the returns from holding nominal bonds and equity over long periods is reflected in the basic model of the paper and that such uncertainty would result in portfolio holders being willing to pay a substantial premium for a long-term indexed bond.

d) Certain institutional factors may make it desirable to introduce index bonds. For example, under present mortgage repayment schedules, the percentage of income (assumed growing at the rate of inflation) devoted to mortgage payments declines more rapidly the greater the rate of inflation. If mortgage payments were indexed, this would no longer be so.[26]

Finally, we note that one of the major issues connected with indexing and wage escalation—the effects of indexing on the stability of the economy—has not been discussed in this paper.

Appendix 1

Continuous-time stochastic processes have been used most extensively in the economics and finance literature by Robert C. Merton. Merton has provided several expositions of the important properties of these processes and also of the continuous-time stochastic dynamic-programming problem (1969, 1971, 1973). This Appendix provides a nonrigorous discussion of the same points.

1. The price dynamics in Section II of the text is assumed to be given by the stochastic differential equation

$$\frac{dP}{P} = \Pi dt + s dz. \tag{1A}$$

The stochastic process described by (1A) is called an Itô process; the stochastic part, *dz*, can be obtained as the limiting process of a suitably defined random walk in discrete time.[27]

---

[26] Franco Modigliani is currently working on a proposal for the introduction of indexed mortgages as a means of avoiding distortions in the time stream of repayments.

[27] See Cox and Miller (1968, pp. 203–10).

The drift of the process, Π, is the expected rate of inflation per unit time. It is defined by

$$\Pi \equiv \lim_{h \to 0} E_t\{[P(t+h) - P(t)]/P(t)\}/h, \tag{2A}$$

where E_t is the expectation operator, conditional on the value of $P(t)$. Similarly, the variance of the process per unit time is defined by

$$s^2 = \lim_{h \to 0} E_t[(\{[P(t+h) - P(t)]/P(t)\} - \Pi h)^2]/h. \tag{3A}$$

2. A discrete time difference equation which satisfies (2A) and (3A) is

$$\frac{P(t+h) - P(t)}{P(t)} = \Pi h + sy(t)\sqrt{h}, \tag{4A}$$

where $y(t)$ is a normal random variable with zero mean and unit variance which is not temporally correlated. The limit as $h \to 0$ of (4A) then describes a Wiener process for the variable $sy(t)\sqrt{h}$, and the equation can be written

$$\begin{aligned}\frac{dP}{P} &= \Pi dt + sy(t)\sqrt{dt} \\ &= \Pi dt + s dz, \qquad \text{where } dz \equiv y(t)\sqrt{dt}.\end{aligned} \tag{5A}$$

3. The important characteristics of the process described by (5A) should be evident from consideration of (4A). First, the increments in dP/P are independent no matter how short the time period. Second, as the time period shrinks, the variance of the stochastic component also shrinks, since the variance is just $s^2 h$. In fact, no jumps in dP/P are possible in the limit, even though dP/P is stochastic. For the process for P described by (5A), an Itô process, P is continuous with probability 1 but it is, with probability 1, not differentiable.

4. In solving the portfolio problem, we deal with a utility function which is a function of several random variables whose behavior can be defined by Itô processes. To handle the problem, Itô's lemma, sometimes called the fundamental theorem of the stochastic calculus, is needed.

Suppose we have a number of stochastic processes describable by

$$\frac{dP_i}{P_i} = \Pi_i dt + s_i dz_i, \qquad i = 1, \ldots, n, \tag{6A}$$

and with ρ_{ij} as the correlation coefficient between the Wiener processes dz_i and dz_j. Then let $F(P_1, \ldots, P_n, t)$ be a function which is at least twice differentiable, which, obviously, depends on the stochastic processes. Itô's lemma gives the rule for finding the differential of $Y = F(P_1, \ldots, P_n, t)$. Specifically,

$$dY = \sum_1^n \frac{\partial F}{\partial P_i} dP_i + \frac{\partial F}{\partial t} dt + \tfrac{1}{2} \sum_i \sum_j \frac{\partial^2 F}{\partial P_i \partial P_j} dP_i dP_j \tag{7A}$$

is the stochastic differential of the function $F(\ )$. The product $dP_i dP_j$ is defined by

$$\begin{aligned} dz_i dz_j &= \rho_{ij} dt, \qquad i, j = 1, \ldots, n, \\ dz_i dt &= 0, \qquad i = 1, \ldots, n. \end{aligned} \tag{8A}$$

5. A simple example may help in understanding Itô's lemma. Consider two Itô processes:

$$\frac{dP}{P} = \Pi dt + s dz, \tag{9A}$$

$$\frac{dQ}{Q} = r_N dt. \tag{10A}$$

We are interested in the stochastic process describing the variable $q \equiv Q/P$. Now

$$\frac{1}{q} d(Q/P) = \frac{1}{q}\left(\frac{1}{P} dQ - \frac{Q}{P^2} dP - \frac{dQdP}{P^2} + \frac{Q}{P^3} dP^2\right), \tag{11A}$$

where the first two terms in parentheses are the first-derivative terms in (7A) and the second two terms in parentheses are the second-derivative terms in (7A). Simplifying (11A), we obtain

$$\begin{aligned}\frac{dq}{q} &= \frac{dQ}{Q} - \frac{dP}{P} - \frac{dQ}{Q}\frac{dP}{P} + \left(\frac{dP}{P}\right)^2 \\ &= (r_N - \Pi + s^2)dt - s dz. \end{aligned}\tag{12A}$$

Itô's lemma is used in recognizing that the term in $(dQ/Q)(dP/P)$ in (12A) disappears and that, in the $(dP/P)^2$ term, we obtain a term $s^2 dz^2$ which is just $s^2 dt$. Note that

$$\frac{dq}{q} \neq \frac{dQ}{Q} - \frac{dP}{P},$$

although one accustomed to working with logarithmic derivatives would be inclined to assume that they are equal.

6. Itô's lemma is also of use in discussing the behavior of the price level implied by (1A). The solution to (1A) is

$$P(t) = P(0) \exp\left[\left(\Pi - \frac{s^2}{2}\right) t + s \int_0^t dz\right]. \tag{13A}$$

(In checking the solution using Itô's lemma, write $P = P(t, z)$ and do not neglect the term in dz^2.) Then, taking logarithms of both sides gives

$$\log \frac{P(t)}{P(0)} = \left(\Pi - \frac{s^2}{2}\right) t + s \int_0^t dz. \tag{14A}$$

Thus, $P(t)$ is log normally distributed. Using (14A) gives

$$E\left[\log \frac{P(t)}{P(0)}\right] = \left(\Pi - \frac{s^2}{2}\right) t, \tag{15A}$$

$$\operatorname{var}\left[\log \frac{P(t)}{P(0)}\right] = s^2 t. \tag{16A}$$

7. Finally, we discuss the continuous-time stochastic dynamic-programming problem.[28] Typically, the problem is to find

$$\max E_0 \int_0^\infty U[C(t), t]\, dt, \tag{17A}$$

---

[28] Dreyfus (1965, chap. 7) provides a lucid derivation of the equation of optimality in the stochastic continuous-time problem.

where $U[C(t), t]$ is strictly concave in the vector $C(t)$. The budget constraint is

$$dW = \sum_{i=1}^{m} w_i(R_i - R)Wdt + \left(RW - \sum_{j=1}^{n} P_jC_j\right) dt + \sum_{i=1}^{m} w_is_iWdz_i, \quad (18A)$$

where w_i are the portfolio shares for risky assets, R the nominal interest rate, P_j the price of the jth consumption good, and R_i and s_i^2 the means and variances of the nominal rates of return on the ith asset.

The discrete time analog would be

$$\max E_0 \sum_{t=0}^{\infty} U[C(t), t],$$

with a budget equation similar to (18) and with C and w_i, the portfolio shares, as the choice variables. To solve the discrete time problem, one writes

ZAK HUS JAMIA NA NEW DELHI-

$$\begin{aligned} J_t(W_t, P_t) &= \max_{\{C, w_i\}} E_t \sum_{t}^{\infty} U[C(t), t] \\ &= U[C^*(t), t] + E_t[J_{t+1}(W_{t+1}, P_{t+1})]. \end{aligned} \quad (20A)$$

At the optimum,

$$\begin{aligned} 0 &= \phi[C^*(t), w^*(t); W_t, P_t, t] \\ &= U[C^*(t), t] + E_t[J_{t+1}(W_{t+1}, P_{t+1})] - J_t(W_t, P_t). \end{aligned}$$

In continuous time, one defines

$$\phi(C, w; W, P, t) = U(C, t) + L(J), \quad (21A)$$

where $L(J)$ is to be thought of as

$$\lim_{h \to 0} E_t \left\{ \frac{J_{t+h}(W_{t+h}, P_{t+h}) - J_t(W_t, P_t)}{h} \right\}$$

for a given set of controls, w and C.

The basic result is given by Merton (1971, p. 381) as theorem 2; it states that, for the P_t generated by Itô processes[29] and the $U(\ )$ concave in C, there exists a set of optimal controls[30] w^* and C^* such that

$$0 = \phi(C^*, w^*; W, P, t) \geq \phi(C, w; W, P, t). \quad (22A)$$

Thus, to find a maximum, one simply maximizes

$$\phi(C, w; W, P, t) = U(C, t) + \lim_{h \to 0} E_t\{[J_{t+h}(W_{t+h}, P_{t+h}) - J_t(W_t, P_t)]/h\} \quad (23A)$$

with respect to C and w. The expectation in (23A) is evaluated by expanding $J_{t+h}(\ )$ in Taylor series around (W_t, P_t). If we write

$$\frac{dW}{W} = \gamma dt + \varepsilon dz, \quad (24A)$$

$$\frac{dP}{P} = y dt + \sigma dz, \quad (25A)$$

---

[29] This is overly strong.

[30] In the infinite-horizon case, a transversality condition has also to be satisfied.

where P is a vector and γ and ε are derived from the budget constraint (18A), we obtain

$$\begin{aligned}\phi(C, w; W, P, t) = {} & U(C, t) + J_t + J_W \gamma W + \tfrac{1}{2} J_{WW} W^2 \varepsilon^2 \\ & + J_P y P + \tfrac{1}{2} J_{PP} \sigma^2 P^2 + J_{WP} \rho_{PW} \sigma \varepsilon P W. \qquad (26A)\end{aligned}$$

In (26A) $J_P yP$ stands for

$$\sum_{i=1}^{n} J_{P_i} y_i P_i,$$

$J_{PP} \sigma^2 P^2$ stands for

$$\sum_{i=1}^{n} \sum_{j=1}^{n} J_{P_i P_j} \sigma_{ij} P_i P_j,$$

8. Once the $\phi(\ )$ function is maximized, one has the differential equation for $J(W, P, t)$, which is now a deterministic equation. Solving this provides the solution to the dynamic-programming problem.

Appendix 2

TABLE OF NOTATION

A. SUBSCRIPTS

Sections II and IV ($i = 1, 2, 3$):
1 = index bond
2 = equity
3 = nominal bond

Section III ($i = 1, \ldots, n + 1$):
$1, \ldots, m$ = goods, and bonds indexed on those goods
$m + 1, \ldots, n$ = equity assets
$n + 1$ = nominal bond

B. ALPHABETICAL

Term	Definition
b^2	Variance of real wage income
dq	Wiener process generating real wage income
r_i	Expected (sometimes with certainty) real rates of return
$s^2 = s_1^2$	Variance of rate of inflation
s_i^2	Variances of nominal returns on assets
w_i	Portfolio shares
$[w]$	$[w_1, \ldots, w_n]'$
dx_i	Wiener processes generating real asset returns
$dz = dz_1$	Wiener process generating inflation rate
dz_i	Wiener processes generating nominal asset returns
$A(\ )$	Inverse of degree of relative risk aversion
C	Rate of consumption
C_i	Rate of consumption of good i
$J(\ )$	Derived utility function
P	Price level
P_i	Prices of goods
R_i	Expected nominal returns on assets
R_3	Nominal interest rate on nominal bond (Sec. II)
$R_{n+1} = R$	Nominal interest rate on nominal bond (Sec. III)
$[\hat{R} - R]$	$[R_1 - R_{n+1}, \ldots, R_n - R_{n+1}]'$
$U(\ )$	Instantaneous utility function
V	Real wealth
W	Nominal wealth
Y	Real wage
δ	Rate of time discount
η	Correlation coefficient of real wage income with rate of inflation
η_2	Correlation coefficient of real wage income with equity return
$\eta_3\ (= -\eta)$	Correlation coefficient of real wage income with real return on nominal bonds
λ	Coefficient of correlation of real returns on equity and nominal bonds
ν	Expected rate of increase of real wage income
Π	Expected rate of inflation
Π_i	Expected rate of increase of price of good i
ρ	Correlation coefficient of nominal returns on equity and index bonds
λ	Correlation coefficient of real returns on equity and the rate of inflation
ρ_{ij}	Correlation coefficient of nominal returns on assets i and j
σ_2^2	Variance of real return on equity
$\sigma_3^2 = s_1^2$	Variance of real return on nominal bonds
Ω	Variance-covariance matrix of nominal returns on assets

References

Ahtiala, Pekka. *Index-linked Debts.* Tampere: Tampereen Yliopisto, 1967.

Cagan, Phillip. "Common Stock Values and Inflation—the Historical Record of Many Countries." National Bureau Report Supplement. New York: Nat. Bur. Econ. Res., March 1974.

Collier, Robert P. *Purchasing Power Bonds and Other Escalated Contracts.* Taipei: Buffalo, 1969.

Cox, D., and Miller, H. *The Theory of Stochastic Processes.* New York: Wiley, 1968.

Dreyfus, Stuart. *Dynamic Programming and the Calculus of Variations.* New York: Academic Press, 1965.

Fama, Eugene F., and MacBeth, James D. "Tests of the Multiperiod Two-Parameter Model." *J. Financial Econ.* 1, no. 1 (May 1974): 43–66.

Fischer, Stanley. "The Demand for Index Bonds." Working Paper no. 132. Cambridge, Mass.: Massachusetts Inst. Tech., Dept. Econ., May 1974.

Fisher, Irving. *Stable Money.* New York: Adelphi, 1934.

Fishlow, Albert. *Indexing Brazilian Style: Inflation without Tears?* Brookings Papers on Economic Activity, no. 1. Washington: Brookings Inst., 1974.

Hanoch, Giora. "Risk Aversion and Consumer Preferences." Discussion Paper no. 366. Cambridge, Mass.: Harvard Inst. Econ. Res., June 1974.

Jones, Robert A. "Price Uncertainty and the Standard of Deferred Payment." Unpublished. Univ. California, Los Angeles, 1974.

Long, John B. "Stock Prices, Inflation, and the Term Structure of Interest Rates." *J. Financial Econ.* 1, no. 2 (July 1974): 131–70.

Merton, Robert C. "Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case." *Rev. Econ. and Statis.* 31, no. 3 (August 1969): 247–57.

———. "Optimum Consumption and Portfolio Rules in a Continuous-Time Model." *J. Econ. Theory* 3, no. 4 (December 1971): 373–413.

———. "An Intertemporal Capital Asset Pricing Model." *Econometrica* 44, no. 5 (September 1973): 867–87.

Modigliani, Franco. "Some Economic Implications of the Indexing of Financial Assets with Special Reference to Mortgages." Unpublished. 1974.

Pigou, A. C. *Industrial Fluctuations.* New York: Macmillan, 1929.

Roll, Richard. "Assets, Money, and Commodity Price Inflation under Uncertainty." *J. Money, Credit, and Banking* 5, no. 4 (November 1973): 903–23.

Samuelson, Paul A., and Swamy, S. "Invariant Economic Index Numbers and Canonical Duality: Survey and Synthesis." *A.E.R.* 64 (September 1974): 566–93.

Sarnat, Marshall. "Purchasing Power Risk, Portfolio Analysis, and the Case for Index-linked Bonds." *J. Money, Credit, and Banking* 5, no. 3 (August 1973): 836–45.

Siegel, Jeremy. "The Theory and Consequences of Contracting in Nominal Terms." Unpublished. Univ. Chicago, 1974.

Tobin, James. "The Theory of Portfolio Selection." In *The Theory of Interest Rates,* edited by F. H. Hahn and F. P. R. Brechling. London: Macmillan, 1965.

———. "An Essay on the Principles of Debt Management." In *Essays in Economics.* Vol. 1. *Macroeconomics.* Chicago: Markham, 1971.

Demography, Technology, and Higher Education: Toward a Formal Model of Educational Adaptation

Stephen P. Dresch

Yale University and National Bureau of Economic Research

The ultimate concern of this paper is educational adaptation, that process by which the educational composition of the adult population is modified in response to changes in technology and economy. Focusing on the changing role of the college educated, the first section descriptively examines the inter- and intrasectoral components of change in educational composition of the labor force, 1929–69, and explicates the destabilizing consequences of demographic movements. The second section develops a more aggregative representation of the demand side of the labor market, utilizing a CES technology relating educated and uneducated labor and incorporating differential rates of factor augmenting technical progress, and conjoins this with an explicit model of the determinants of the educational composition of the adult population. In the context of future demographic movements, the model is employed in the third section to project a 33 percent decline in college enrollments between 1970 and 2000. The final section then places the simulated educational adaptation process in its larger social-demographic context.

Between 1955 and 1963 undergraduate college and university enrollments increased at an average annual rate of 6.1 percent. For the period 1963–70 this rate reached 7.9 percent. Reflecting not only a fascination with growth per se but also the attraction of the real rewards of growth, it is not surprising that myopic higher educators, given such dramatic

I would acknowledge the somewhat unintended stimulus preparation of this paper received from Hank Lieberman of the *New York Times* and Earl Cheit of the Carnegie Council (then of the Ford Foundation); the contributions to my thinking of conversations with Dietrich Goldschmidt, Richard Nelson, and Derek Price; and the participation and assistance of Adair Waldenberg and Nellie Zorc. The charts have been prepared by H. Irving Forman with his customary skill. Preparation of this paper has been supported financially by a grant to Yale University from the Alfred P. Sloan Foundation. The study has not undergone review by the Board of Directors of the National Bureau of Economic Research and is therefore not an official publication of that organization.
[*Journal of Political Economy*, 1975, vol. 83, no. 3]
© 1975 by The University of Chicago. All rights reserved.

stimulae, converted a temporary, highly unique situation into a presumed permanent state of affairs.

As early as the mid-1960s, however, the first shadows began to pass over this bright vision of the future, as Allan Cartter (1965) and others pointed out a basic demographic fact of life: a necessary, if not sufficient, condition for the phenomenal growth of college enrollment in the late 1950s and 1960s was the rapid growth of the college-age population, stemming from the war and postwar explosion of births following the depressed birth rates of the 1930s. Clearly, a major source of the observed rate of growth in higher education between 1955 and 1970 was demographic. In 1960, 273,000 24-year-olds had completed college, a figure which had increased by 1972 to 913,000, an increase of 235 percent. Over the same period, however, the *proportion* of 24-year-olds who had completed college increased only from 13 percent to 26 percent, or by not more than 100 percent. Thus, while the net spread by higher education was widened somewhat, the number of persons actually snared increased much more than proportionately.

Cartter's critique of the unlimited growth perspective, emphasizing the implications of the declining birth rates of the 1960s for the size of college-age cohorts in the 1980s, brought into graphic focus the fact that there existed very definite demographic constraints on growth, that the world of the late twentieth century was not one of constant uniform population growth but one of contraction. And unlike earlier stagnationist theses, which were concerned with prospective declines in population, these demographic observations were given added weight by the fact that all of the persons of relevance to the analysis had already been born.

In the wake of Cartter's revelations trend-demographic (*t-d*) enrollment projections have become quite popular. In contrast to earlier manpower-requirements approaches to educational planning and projection, which assume a rigid compositional structure of the labor force, the *t-d* approach ignores entirely the relation of education to production and to the structure of the labor force.[1]

Representative examples of *t-d* projections are provided by the federal government's National Center for Educational Statistics (NCES 1972, 1973) and by the Carnegie Commission on the Future of Higher Education (1973). Oversimplifying slightly, NCES projects the *ratio* of enrollment to college-age population, based on data for preceding years, and then applies the projected ratio to estimates of the future size of the college-age population to obtain projections of enrollment. Projections of educational attainments (e.g., degrees awarded) are then derived via similar techniques from projections of enrollment in immediately preceding years.

---

[1] For an example of the manpower-planning methodology, see Parnes (1962).

Thus, the enrollment ratios provide the core of the NCES projections. Because the ratio of enrollment to eligible population, y, cannot exceed unity, an upper limit, K, is selected, "based on consideration of the statistic being projected" (NCES 1973, p. 119). A logistic function is then fitted statistically, employing a transformation of the equation $y = K[1 + e^{-(a+bt)}]$.[2] While the Carnegie Commission's projection methods are not thoroughly described, they seem to be similar, if somewhat more general (incorporating probable changes in educational options, proliferating adult programs, etc.).[3] Examples of the resultant projections of enrollment change, by decade, are provided in table 1.

The arbitrary nature of these projections is indicated by the significant year-to-year shifts. Thus, the 1973 Carnegie Projection II predicts an increase in enrollment (1970–80) only approximately 60 percent as great as the 1971 Carnegie Projection I, and a decline (1980–90) twice as great, due solely to the inclusion of later data. Similarly, the 1972 NCES projection of the increase in enrollment, 1970–80, is significantly less than the 1971 NCES projection for the same reason, and preliminary evidence suggests that the same downward trend will be followed by the 1973 NCES projection (NCES 1973, p. iv).

The fundamental limitation of these analyses resides in their consideration only of the demographic side of the system, that is, of prospective changes in the size of the college-age population, assuming a simple perpetuation of past enrollment trends. They explicitly ignore (*a*) the capacity of the broader system (economy) to absorb the college educated, for example, the patterns of change in the interindustry composition of employment (a factor overemphasized by the manpower planners); and (*b*) the effect on the overall supply of highly educated persons of the rapid growth in the college-age population from the late 1950s to the mid-1970s, especially considering the fact that this bulge followed a decline in the college-age population caused by the extremely low birth rates of the 1930s. Thus, the significance of the absolute decline in the college-age population after 1980 resides not only in the decline itself but in the fact that it succeeds a period of rapid population growth and of rapidly increasing rates of college attendance, the latter a consequence of preceding economic and demographic phenomena. If the capacity of the system to absorb the college educated is limited, then it may not be possible in the future even to sustain current rates of college attendance, let alone levels of enrollment.

The inadequacies of the *t-d* approach, it must be noted, have been permitted and to a degree justified by the emphasis in the human-capital

---

[2] The transformation is simply $z = a + bt$ where $z = \ln[y/(K - y)]$ (see NCES 1973, pp. 119–20).

[3] The Carnegie Commission's enrollment projections are described in Technical Note A (1973, pp. 95–106).

TABLE 1

PROJECTIONS OF PERCENTAGE CHANGE IN COLLEGE-AGE POPULATION AND PREBACCALAUREATE ENROLLMENT, BY DECADE 1970–2000

DECADE	NCES* (1971)	NCES* (1972)	CARNEGIE COMMISSION† Projection I (1971)	Projection II (1973)	ED. ADAPT. MODEL QUASI PROJECTION‡	POPULATION§ 18–24
1970–80	54.7	43.2	48.9	30.6	40.9	19.2
1980–90	...	...	−4.5	−8.6	−45.8	−13.9
1990–2000	...	...	33.4	26.3	−11.9	17.8
1970–2000	...	...	89.7	50.8	−32.7	21.0

* Total enrollment, including nondegree credit, from NCES, 1972 and 1973, respectively.

† Prebaccalaureate, including nondegree credit.

‡ Estimates of degree-credit enrollment only, based on projections of numbers of 24-year-olds with baccalaureate degrees in 1974, 1984, ...

§ Census Series E projections.

literature on the estimation of static rates of return to investments in education, a preoccupation criticized by Welch (1970), and by empirical evidence that relative wages (and rates of return) are independent of relative endowments of different types of labor (Bowles 1970; Griliches 1970; Psacharopoulos and Hinchliffe 1972). The conclusion of the present study, however, is that the apparent capacity of the economy to absorb highly educated labor without depressing relative wages has been the product of a unique succession of economic-technological and demographic events and that the historical consequences of these events are unlikely to be replicated over the intermediate future.

The first section places the development of higher education in historical perspective, focusing on the economic and demographic profile of the period 1930–70. A general model of educational adaptation is then developed in the second section and applied in the third section to project movements in college-level educational attainments over the next 30 years. The final section then explores several important concomitants and consequences of the anticipated process of educational adaptation.

I. Economic and Demographic Underpinnings of the Growth of Higher Education, 1930–70

The significant external forces underlying the growth of higher education over the course of the twentieth century can be grouped under two headings: population dynamics and intersectoral shifts in the composition of economic activity. Consider first the radical changes in industrial composition, focusing on the period 1929–69 (years for which detailed data are available).[4]

Changes in the Composition of Economic Activity, 1929–69

Table 2 presents growth rates and employment shares by sector for the full period 1929–69 and for the subperiods 1929–48 and 1948–69. Growth in total employment over the period, at an average rate of 1.4 percent per annum, was relatively evenly distributed between the subperiods, with an average annual rate of 1.3 percent in the first and 1.5 percent in the second. By contrast, the rate of change in the sectoral composition of employment, although significant in both periods, was much greater in the second, as the differential subperiod growth rates of the rapidly expanding (and contracting) sectors indicate. Thus, over the entire period the eleven most rapidly expanding sectors (defined as those with

---

[4] All data relating to the industrial composition of employment in the years 1929, 1948, and 1969 are derived from Denison (1973, table A-1, p. 35).

TABLE 2

Employment Change by Sector, 1929–69, and College-educated Share of Labor Force, 1970

	Mean Annual Growth Rate (%)			Share of Total Emp.			College Educated in 1970 (%)*
	1929–48	1948–69	1929–69	1929	1948	1969	
All industries	1.3	1.5	1.4	1.0	1.0	1.0	12.9
Farms	−1.5	−3.6	−2.6	.194	.114	.039	2.9
Agricultural services, et al.	0.8	0.8	0.8	.005	.005	.004	8.2
Coal mining	−0.7	−6.3	−3.7	.014	.009	.002	1.7
Other mining	1.2	0.2	.7	.008	.008	.006	12.0
Construction	1.8	1.3	1.6	.050	.055	.054	3.9
Food, kindred products	2.8	−0.1	1.3	.023	.031	.023	5.1
Textile-mill products	0.3	−1.3	−0.6	.027	.023	.013	3.4
Other nondurable manufacturing	1.8	1.4	1.6	.065	.071	.070	8.5
Durable manufacturing	2.5	1.6	2.1	.116	.146	.152	7.9
Transportation	−0.0	−0.6	−0.3	.066	.051	.033	4.2
Telephone and telegraph	1.4	1.4	1.4	.012	.012	.012	5.2
Radio, TV broadcasting	13.1	4.4	8.5	.000	.001	.001	21.7
Electric, gas, sanitary	0.4	1.2	0.8	.011	.009	.009	8.3
Wholesale trade	2.2	1.6	1.9	.038	.045	.047	9.9
Retail trade and auto services	1.7	2.7	2.2	.131	.143	.186	4.4
Banking	0.3	4.1	2.3	.008	.007	.012	12.2
Other finance, insurance, real estate	1.3	2.6	2.0	.026	.026	.033	17.4
Hotels and lodging places	1.1	1.1	1.1	.011	.011	.010	4.2
Personal services	1.1	0.8	0.9	.022	.021	.018	2.2
Misc. business and profit services	4.3	6.1	5.3	.006	.011	.030	28.6
Misc. repair services	3.5	1.0	2.2	.003	.004	.004	5.5
Motion pictures	2.2	−0.9	0.6	.003	.004	.002	13.3
Other amusements	0.1	2.5	1.3	.006	.005	.006	6.2
Medical and other health	2.2	4.9	3.6	.016	.019	.040	18.6
Legal services	0.6	2.7	1.7	.004	.004	.005	52.1
Educational services	2.3	4.5	3.5	.007	.008	.016	42.4
Nonprofit organizations	3.2	3.5	3.4	.008	.011	.017	16.3
Private households	−2.1	−0.6	−1.3	.051	.027	.017	1.0
Government, gov't ent.	4.0	3.6	3.8	.069	.116	.181	34.5
Estimated actual college educated (%)				5.2	6.7	12.9	
Hypothetical college educated (%)*				7.5	9.5	12.9	

Sources: Denison 1973 and U.S. Bureau of the Census 1973, pp. 11–22.

* Obtained by applying percentage college educated by sector in 1970 to sectoral distributions of employment in 1969, 1948, and 1929.

mean annual growth rates in excess of 2 percent) increased their share of total employment from 39.0 to 67.2 percent; but of this 28.2 percentage point gain, 10.2 points were added during the first period, 18.0 during the second.

Only a casual review of the most rapidly expanding sectors is necessary to grasp the significance of their expansion for the higher-education sector. While estimates by sector of the college-educated proportion of the labor force are available only for 1970, these data indicate that of the eleven rapid-growth sectors, four exhibited college-educated proportions less than the mean for the entire labor force; but, of the five sectors in which the college educated comprised over 20 percent of the labor force, four were rapid growth. On the whole, the most rapidly expanding sectors were clearly the most intensive employers of the college educated.

Admittedly, educational attainments of personnel in most sectors have probably increased over this period. Changes in products, the introduction of new products, changes in production processes, and changes in the nature and complexity of economic organization all combine to constitute a pervasive increase in technological (including organizational) sophistication, explaining at least in part the substantial increases in the educational attainments of the labor force. However, as indicated by table 3, the intersectoral shift in output and employment which has occurred over this 40-year span accounts for most of the change in the college-educated share of the labor force.

Between 1929 and 1969 the college-educated proportion of the labor force increased from 5.2 to 12.9 percent, a gain of 7.7 percentage points. To assess the effect of the change in the interindustry composition of employment, the 1970 estimates of college-educated proportions by sector were applied to the 1929 and 1948 distributions of employment by sector, generating hypothetical educated (meaning college educated) shares of the labor force in these years.

Of the 7.7-point gain in the educated share over the full period, 5.4 percentage points are accounted for by the change in sectoral composition. Only a residual 2.3 percentage points can be explained by general intrasectoral increases in relative employment of the college educated. Thus, of the overall change, 70 percent can be allocated to intersectoral shifts and only 30 percent to residual increases in educational attainments within sectors. Clearly, for the period as a whole the dominant factor underlying the increase in the relative representation of the college educated is the rapid growth of education-intensive industries.

The full significance of this change, however, derives from its very peculiar timing. Rather than evolving smoothly over the entire 40-year span, over 80 percent of the total increase in college-level educational attainments occurs after 1948. That the postwar period was one of more

TABLE 3

INTER- AND INTRASECTORAL COMPONENTS OF GROWTH IN COLLEGE-EDUCATED SHARE OF LABOR FORCE, 1929–48–69

A. PERCENTAGE OF LABOR FORCE COLLEGE EDUCATED

Year	Actual	Hypothetical*	Ratio of Actual to Hypothetical
1929	5.2	7.5	.70
1948	6.7	9.5	.70
1969	12.9	12.9	1.0

B. COMPONENTS OF CHANGE

	Percentage Point	Percent of Total Change	Percent of Component
For 1929–48	1.5	19	...
For 1948–69	6.2	81	...
Total 1929–69 ...	7.7	100	
Allocable to interindustry employment shift:*			
1929–48	2.0	26	37
1948–69	3.4	44	63
Total	5.4	70	100
Residual allocable to intraindustry change in percent educated:			
1929–48	−0.5	−6	−22
1948–69	2.8	36	122
Total	2.3	30	100

SOURCE.—Derived from table 2.
* Assume college educated proportions *by sector* in 1929, 1948, and 1969 equal to observed 1970 values.

rapid change in economic structure is hardly surprising. Between 1929 and 1948 a succession of major events intervened to prevent the apparent continuation and realization of the economic and technological tendencies which had become operative earlier. For example, the depression of the 1930s constituted a sharp disruption in the growth of the service sectors and in the rapid decline of such traditional sectors as agriculture and mining. Similarly, the economic impact of the war was to temporarily reverse the trend away from primary raw materials and manufacturing industries. Thus, it is only in the postwar expansion that the accumulating changes in economic and technological structures could be effected.

As a result, of the 5.4 percentage-point increase in college-level educational attainments due to intersectoral shifts in employment over the full period, only 2.0 points (37 percent) occur prior to 1948, as against 3.4 points (63 percent) between 1948 and 1969. Thus, the increase in demand for college-educated labor due to changes in the sectoral composition of employment is overwhelmingly a phenomenon of the postwar period.

Notwithstanding this concentration of the intersectoral shift effect in the post-1948 period, the effect of intrasectoral increases in educational attainments is even more extremely concentrated in the postwar period. Specifically, the sectoral representation of the college educated did not undergo any general change between 1929 and 1948; in both years actual educated shares of the labor force were just equal to 70 percent of the hypothetical educated shares (obtained by applying 1970 proportions of the college educated to actual sectoral distributions of employment).

Thus, if for the sake of argument it is assumed that the hypothetical educated shares in 1929 and 1948 represented employer demands at prevailing wage differentials, then the gap between actual and hypothetical educated shares, representing an excess demand for the college educated, widened by 0.5 percentage points between 1929 and 1948, a deficit in college-educated personnel equal to −6 percent of the 7.7 percentage-point total increase in the college-educated share between 1929 and 1969, or −22 percent of the total 2.3-point increase due to the intrasectoral increases in educational intensities. In short, between 1929 and 1948 the educated share of the labor force actually lost ground relative to the overall gain between 1929 and 1970.

It might be argued that this failure of general levels of educational attainment to increase between 1929 and 1948 was primarily a product of the depression. Thus, the high rates of unemployment in the 1930s reflected a temporary excess supply of all types of labor and might have been expected to have greatly reduced the incentive for (and ability of) young entrants into the labor force to obtain college educations. Actually, this does not seem to have occurred. The average annual rate of growth enrollment did in fact decline from 6.1 percent during the 1920s to 3.1 percent during the 1930s, but this is primarily explained by an even more substantial decline in the growth rate of college-age cohorts (18–24) from 1.7 to 0.7 percent. As a result, enrollment as a proportion of the age cohort increased from 7.2 percent in 1930 to 9.1 percent in 1940. The most important factor explaining the failure of college enrollments to decline over the course of the depression is the impact of high rates of unemployment on the foregone earnings cost of college attendance.[5] If anything, labor-market conditions served to perpetuate school attendance.

In any event, had the depression not retarded the intersectoral developments which provided the primary basis for the increases in the educated share of the labor force observed over the full period, then enrollment would have increased at a much more rapid rate between 1930 and 1940. Given the decline in the growth rate of college-age cohorts, a more substantial increase in the proportion enrolled in college would have taken place than was actually observed. The observed increase in the proportion enrolled was just consistent with the limited change in the sectoral composition of employment which did take place over the decade of the 1930s.

---

[5] This interpretation is supported by Crean's (1973) evidence on cyclical movements in Canadian secondary schools' enrollment rates.

The significant point is that the depression, the trend during the 1930s in the size of college-age cohorts (due to changes in immigration laws and to the low birth rates during and after World War I), and World War II together produced a discontinuity in the broad trend of developments between 1929 and 1969 which could not be easily and quickly accommodated in the period of immediate postwar readjustment.

Destabilizing Consequences of Population Dynamics

The postwar adjustments were rendered particularly difficult by the radical decline in completed fertility rates following World War I compounded by the depression-induced demographic disruptions of the 1930s. While completed fertility rates had followed a somewhat erratic but declining path prior to 1920, the remarkable decline in births during the 1920s and 1930s led to an absolute contraction in the size of college-age cohorts after World War II. The growth rate of the 18–24-year-old age group, which had averaged above 1 percent between 1900 and 1940, declined to −0.2 percent in 1940–50 and −0.3 percent in 1950–60. Thus, just as the economy was able to incorporate the technological and organizational changes which had been held in abeyance by depression and war, the age group for which educational attainments were subject to modification began to contract significantly. Although college enrollment relative to the 18–24-year-old population experienced radical increases between 1940 and 1960 (from 9.1 to 20.5 percent), the decade of the 1950s witnessed the lowest average rate of enrollment growth in this century, 1.9 percent. Thus, between 1940 and 1960, while the proportion of young people enrolled in college increased by 125 percent, the college-educated proportion of the adult population increased only by 67 percent, from 4.6 to 7.7 percent.

In this way, the dramatic increase in enrollment rates can be seen as a consequence of the sudden effectuation of a previously unobserved change in labor-market demand submerged by depression and war on the one hand, and a marked decline in the size of educationally malleable age cohorts, similarly at least partially a product of the events of the 1930–45 period, on the other. The full consequences of this increase in the college attendance rate, however, depended also on the demographic events which followed the war: the postwar baby boom and the continuing high fertility rates of the 1950s. After declining at an annual rate of −0.3 percent between 1950 and 1960, the 18–24-age group increased at a rate of 4.5 percent per annum from 1960 to 1970. Correspondingly, the rate of growth of college enrollment increased from 1.9 to 8.5 percent, while the proportion of the age cohort enrolled increased from 20.5 to 30.6 percent.

Thus, the postwar period has been one in which previously repressed economic and technological forces became effective precisely at a time when demographic factors were least conducive to significant changes in the overall educational composition of the adult population, generating substantial increases in the rate at which the contracting educationally malleable age cohorts were in fact educated. The momentum of this movement toward higher college attendance, however, reached its peak just as the emaciated college-age cohorts born in the 1930s were replaced by the swollen cohorts of the postwar baby boom.

Effectively, after the war the excess demand for college-educated *persons* drove up the *proportions* of the relevant age cohorts educated. But, as will be discussed further below, it is this proportion which serves as a critical determinant of the rate at which later cohorts are educated. Thus, as the size of these cohorts suddenly and discontinuously increased after 1964 (reflecting the first major bulge in births in 1946), the inertia in the system, operating through the proportions educated, forced up disproportionately the *number* of persons educated.

As a result of the rapid increase in the size of college-age cohorts relative to the adult (over age 25) population and of the concomitant rise in the rate at which these cohorts were educated, the college-educated proportion of the total adult population, which had increased gently from 4.6 percent in 1940 to 6.0 percent in 1950 and 7.7 percent in 1960 (i.e., by between 1.4 and 1.7 percentage points per decade), quickly climbed to 11.0 percent in 1970 (an increase of 3.3 percentage points over the decade) and further to 12.0 percent in 1972. Subject to very little prediction error, a further increase to 15 or 16 percent by 1980 can be anticipated, for a total gain of 4 or 5 percentage points over the decade 1970–80.

Thus, the inertia generated in the rapid expansion of the period approximately 1950–65 carries the system forward at an accelerating rate between 1965 and 1975, due to the discontinuous increase in the size of college-age cohorts. In its origins the expansion process reflected a suddenly operative excess demand for the college educated, while in its latter phases (after 1965) the dominant factor is the rapid expansion in the supply of educationally malleable labor, the educational attainments of which are determined by an excess demand existing in the past but quickly being eliminated.

II. A Model of Educational Adaptation

In Section I the change in college-level educational attainments of the labor force, 1929–69, was interpreted in terms of changes in the interindustry composition of employment and of increases in the representation of the college educated within industries as a result of general increases in technological sophistication. Changes in enrollment rates (ratios of enrollment to college-age population) were then interpreted in terms of these changes in what can now be characterized as the demand for labor, on the one hand, and of the particular demographic profile of the period (i.e., the wavelike movement in the college-age population), on the other. For purposes of projecting future changes in the college-level educational attainments, a more aggregative, composite representation of the demand side of the labor market is developed and conjoined with an explicit representation of the determinants of enrollment rates in the context of a predetermined demographic profile. In the following section the model is employed, in the context of historically determined demographic developments, to simulate the educational attainments of entrants into the adult population and of the entire adult population over the period 1960–2005.

In broad outline the primary focus of the model is on the educational characteristics of the active adult population (AAP), defined as the population between the ages of 25 and 64. No distinction is made between the AAP and the labor force; it is implicitly assumed that persons not in the labor force are engaged in comparably productive nonmarket activity.

At any point in time the educational characteristics of the AAP are fixed. For purposes of simplification, educated members of the AAP are defined as those who have completed 4 or more years of college. All others are classified as uneducated. A somewhat oversimplifying assumption is that the educational characteristics of the AAP can be altered in only two ways: (1) through the educational attainments of entering (25-year-old) cohorts, and (2) through the educational characteristics of persons leaving the AAP (through death, which follows the pattern of age-specific mortality, or through retirement, which occurs automatically at age 65). Thus, the educational attainments of any member of the AAP are fixed at entrance (i.e., at age 25), with no change in educational characteristics permitted after age 25.

The second critical assumption of the model concerns the wage of educated relative to uneducated labor: *at any point in time* the relative educated wage is assumed to be inversely related (although not necessarily inversely proportionate) to the ratio of educated to uneducated members of the AAP. Further, the relative wage is assumed to adjust instantaneously to any change in the educational composition of the AAP.

Over time, however, technological change is assumed to occur, altering the relative educated wage for any division of the AAP between the educated and the uneducated. Thus, at any moment an increase in the relative number of educated persons reduces the relative wage of educated labor, but over time the process of technological change may be such as to permit a rising ratio of educated to uneducated labor without a corresponding reduction in the educated wage differential. Technological change in this context must be interpreted to incorporate both changes in technology within individual sectors and also redistributions of employment over sectors characterized by technologies utilizing different mixes of educated and uneducated labor.

The final critical assumption is that the educational characteristics of entering cohorts (i.e., of 25-year-olds) are determined by educational wage differentials which they observed in prior periods, specifically between the ages of 17 and 24. Any increase in the educated wage differential at one point in time will increase the educated proportions of immediately following cohorts of 25-year-olds.

In effect, then, the model determines the current wage differential on the basis of relative endowments of educated and uneducated labor (members of the AAP). Given current and past wage relatives, the educational characteristics of the next entering cohort are determined. Also, persons leaving the AAP, characterized by age and educational attainment, are identified. Thus, the educational characteristics of the AAP in each successive period are determined. Repeating the process, the model propels itself into the future.

At the outset, it should be emphasized that the model developed here serves primarily a heuristic function, indicating in broad outline the general nature of the process which must be captured by any meaningfully developed representation of the determinants of educational attainment. Because of the inadequacy of available data, at best the model can be implemented only on the basis of crude estimates of important parameters, and in many cases these cannot be estimated at all. Thus, the resultant simulation is, as Martin Shubik has characterized all simulations, quite literally a "fake". That is, it does not purport to predict the future per se, but to indicate explicitly the nature of the process by which the future evolves out of the past and present.

Relative Wage Determination

The process of relative wage determination incorporated in the model assumes a perfectly competitive labor market and a productive technology which identifies two distinct categories of labor, educated (E) and uneducated (U).[6] Specifically, assume that Y, aggregate output (national income, including the value of output and income generated in nonmarket activities),[7] is a function of composite labor (L)[8] and other factors (K), with a production function of the form

$$Y = f(K, L), \text{ where } L = L(E, U). \tag{1}$$

Assuming constant returns to scale (i.e., that the functions f and L are homogenous of degree 1) and a competitive labor market, the absolute wage of educated (uneducated) labor, $W_e(W_u)$, is equal to its marginal product:

$$W_e = f_L(\partial L/\partial E), \tag{2.1}$$

$$W_u = f_L(\partial L/\partial U). \tag{2.2}$$

Thus, the relative wage of educated labor is simply

$$w = W_e/W_u = (\partial L/\partial E)/(\partial L/\partial U). \tag{3}$$

[6] Earlier variants of the type of wage-determination model developed here have been proposed by Bowles (1970), Griliches (1970), Dougherty (1972), and Psacharopoulos and Hinchliffe (1972). A critique of these studies and a recasting of the labor aggregation problem in an explicit supply-demand context is provided by Tinbergen (1974) and, in a somewhat more sophisticated variant, by Fallon and Layard (1975). Neither of these latter studies, however, takes into account the demographic constraints on educational adaptation, and the supply facets are clearly rudimentary.

[7] Specifically included in the conceptual definition of output is net investment in human capital, in the form, e.g., of on-the-job training and learning by doing. Empirically this is incorporated by treating the wage of any type of labor as an annuity paid over the working lifetime, the equal installments of which have the same present value as the predicted stream of lifetime earnings. This adjusts for the differently shaped lifetime-earnings profiles and changing age distributions of educated and uneducated labor.

[8] As will become evident, it is at this stage a matter of indifference whether educated and uneducated labor enter directly into the aggregate production function or through the composite L. Use of the latter assumption simply permits us to ignore the precise characteristics of the entire function f, emphasizing instead the relationships between U and E. By reference to the preceding footnote it can be seen that this procedure is not entirely legitimate, since the discount rate employed in deriving the wage of any category of labor should be related to the rate of return to capital. It is implicitly assumed that the capital stock adjusts to maintain a constant real net rate of return, specifically 3 percent. The possibility of capital-educated labor complementarity (Griliches 1969, 1970; Welch 1970; Psacharopoulos and Hinchliffe 1972) would argue for explicit incorporation of the underlying production function. However, as pointed out by Dougherty (1972), collinearity between capital-labor ratios and time would preclude any meaningful estimates, even given suitable capital-stock data. More seriously, the shortcut of including the capital-labor ratio as a shift factor in a relative wage equation imposes very specific assumptions concerning the characteristics of the aggregate production function. Fallon and Layard (1975) do develop capital-stock data for an international cross-section study and conclude that the existence of capital-educated labor complementarity accounts for the high elasticities of substitution between educationally distinct groups found by Bowles, Dougherty, and Psacharopoulos and Hinchliffe; their corrected elasticity is significantly lower. A similar argument and similar findings are presented by Tinbergen (1974).

One of the most general, and flexible, forms of the composite labor function assumes a constant elasticity of substitution (CES) between educated and uneducated labor,

$$L = (\beta U^{-\rho} + \gamma E^{-\rho})^{-1/\rho}, \tag{4}$$

where $\rho \geq -1$, resulting in the absolute and relative educated wage relationships

$$W_e = f_L \gamma E^{-\rho-1} H, \tag{2.1a}$$

$$W_u = f_L \beta U^{-\rho-1} H, \tag{2.2a}$$

$$w = (\gamma/\beta)(U/E)^{1+\rho}, \tag{3a}$$

where $H = (\beta U^{-\rho} + \gamma E^{-\rho})^{-(1+\rho)/\rho}$.

The most attractive feature of the CES form of the composite labor function is that it permits any degree of substitutability between educated and uneducated labor. Specifically, the elasticity of substitution between U and E, defined as $\sigma = 1/(1 + \rho)$, has limiting values of (a) zero as ρ goes to infinity and (*b*) infinity as ρ goes to -1. In the first case educated and uneducated labor are complementary (i.e., required in fixed proportions). In the second ($\rho = -1$), educated labor simply represents a different number of efficiency units of labor than uneducated labor, and the relative wage is independent of relative endowments of the two types of labor (i.e., education is purely labor augmenting). Because a priori one would expect that substitution possibilities between two categories of the same general factor would be quite high, an expectation supported by previous studies (Bowles 1970; Dougherty 1972), we anticipate a value of ρ in the range -1 to 0.[9]

Technological Change and Secular Modification of the Relative Educated Wage Relationship

The most direct means of introducing technological change into a model of the foregoing type is to assume factor-augmenting progress, that is, that over time the individual labor inputs become more efficient. Thus, the composite labor function, equation (4), becomes

$$L = [\beta(e^{g\tau}U)^{-\rho} + \gamma(e^{h\tau}E)^{-\rho}]^{-1/\rho}, \tag{4a}$$

where τ refers to time, g is the rate of augmentation of uneducated labor, and h is the rate of augmentation of educated labor. However, while factor-augmenting progress operating on either E or U increases the absolute wages of both educated and uneducated labor, its effect on relative wages (i.e., on the share of output accruing to either factor) depends on the relative rates of factor augmentation and on the value of the parameter ρ (i.e., on the elasticity of substitution between educated and uneducated labor $[1/(1 + \rho)]$):

---

[9] Clearly, the CES function, equation (4), is not defined for a value of $\rho = 0$. However, it can be shown that the derived relative wage relationship in this case corresponds to a Cobb-Douglas technology, $L = U^{\beta}E^{\gamma}$, where $\beta + \gamma = 1$, with unitary elasticity of substitution.

$$w = (\gamma/\beta)\{\exp[(g - h)\rho\tau]\}(U/E)^{1+\rho}. \tag{3b}$$

If $\rho = 0$ (the composite labor function is Cobb-Douglas, with unitary elasticity of substitution), the relative shares of educated and uneducated labor are invariant with respect to the rates of augmentation; the augmentation of either factor is just offset by a decline in its wage per efficiency unit, leaving the relative wage per natural unit (person) unchanged. If $\rho \neq 0$, four distinct possibilities can be identified.

1. $g < h$, $\rho < 0$. The rate of augmentation of *educated* labor (h) is greater than that of uneducated labor (g), and the elasticity of substitution $[1/(1 + \rho)]$ is greater than one (i.e., the share of educated labor *increases*). Thus, the educated wage differential at any (U/E) increases over time.
2. $g > h$, $\rho > 0$. The rate of augmentation of *uneducated* labor is greater than that of educated labor and the elasticity of substitution is less than one. The more rapid rate of increase in efficiency units of uneducated labor results in a decline in its *share* and an increase in the relative educated wage at any (U/E).
3. $g < h$, $\rho > 0$. The efficiency of educated labor increases more rapidly but its share of output declines (the elasticity of substitution is less than one), reducing the educated wage relative.
4. $g > h$, $\rho < 0$. The efficiency of uneducated labor increases more rapidly and the share of uneducated labor rises (the elasticity of substitution is greater than one), reducing the educated wage differential.

In fact, as will be discussed further, the ratio of educated to uneducated labor has increased greatly over the last half-century and yet the educated wage differential has not declined substantially, suggesting that the cases of interest are the first two. The anticipation of an elasticity of substitution between educated and uneducated labor significantly greater than unity further narrows attention to the first case.

Unfortunately, empirical estimation of the parameters of equation (3b) is severely constrained by the inadequacy of available data. Wage estimates are based on the present value of cumulative income between the ages of 25 and 64, for high school graduates (uneducated) and for college graduates (educated); the estimates of relative wages and of the educated and uneducated populations are derived from the Current Population Surveys and can be obtained for only 8 years over the period

1956–68.[10] This seriously limits our ability to estimate long-term patterns of change. However, parameter estimates obtained even on the basis of such inadequate data, if consistent with a priori expectations, can at least serve to suggest the nature of the process which is taking place.

For purposes of estimation equation (3b) was multiplied by E/U, converting from relative wages to relative shares, and the following estimates were obtained (t-ratios in parentheses):[11]

$$\begin{aligned} \ln (W_e E/W_u U) &= \ln (\gamma/\beta) + [(g - h)\rho]t + \rho \ln (U/E) \\ &= \underset{(0.22)}{-0.163 + 0.006t} - \underset{(-0.89)}{0.738} \ln (U/E),\ R^2 = .97, \end{aligned}$$

resulting in the basic parameter estimates $\rho = -0.738$, $\gamma/\beta = 0.850$, $(g - h) = -0.008$, $\sigma_{E,U} = 3.819$.

Consistent with our prior expectations, the estimated equation falls into the first class, with more rapid augmentation of educated labor (greater by 0.8 percent per year than the augmentation rate of uneducated labor) and an elasticity of substitution (3.8) significantly greater than unity. A very high elasticity of substitution between two similar factors was anticipated, and the observed signs of ρ and $(g - h)$ are consistent with the hypothesis that technology has been the source of the virtual constancy of the relative educated wage over a period in which the college-educated proportion of the AAP has increased significantly.

The evidence presented in Section I suggests that the estimated pattern of technological change is sensible. Specifically, a rather casual review of the last century of technological (and organizational) development suggests that a secularly more rapid rate of augmentation of educated labor is in fact a reasonable characterization of modern economic history. Increased complexity, and an acceleration in the rate of change, of technology (which in its origins may, or may not, be related to general

---

[10] Data utilized in the estimates of the CES composite labor function parameters were derived as follows: W_e and W_u, respectively, were computed on the basis of expected lifetime incomes (in constant dollars) of 25-year-old males (*a*) who had completed 4 or more years of college (educated), and (*b*) who had completed only high school (uneducated), assuming a 3 percent real discount rate and a 2 percent annual increase in productivity. The underlying data can be found in U.S. Bureau of the Census, Series P-60, pp. 38–39, 46–47, 50–51, 54–55, 58–59, 62–63, 66–67, 70–71. The ratio of educated to uneducated labor (U/E) was computed from U.S. Bureau of the Census, Series P-20, pp. 18, 9, 13, 7, 9, 18, 7, 8, and 9 of nos. 45, 77, 99, 121, 138, 158, 169, 172, and 182, respectively.

[11] These estimates clearly suffer from the collinearity between τ and (U/E), as indicated by the t-values. However, confidence in the estimate is increased somewhat by the fact that the estimate of ρ is not substantially different in an equation which excludes time:

$$\ln(W_e E/W_u U) = 0.287 - \underset{(-12.97)}{0.921} \ln(U/E),\ R^2 = .97.$$

increases in educational attainments), concomitant increases in organizational size and complexity (increasing the requirements for effective coordination), and secular change in the sectoral composition of output and employment would be expected to have widened (strictly speaking, deepened) the role of highly trained personnel, as reflected in an increasing ratio of educated to uneducated labor.[12]

Determinants of Educational Attainments of Entrants into the AAP

For purposes of simplicity the model assumes that the educational attainments of current members of the AAP are completely fixed. Thus, the educated share of the AAP can change over time only as a result of differential educational attainments of new entrants relative to those of persons exiting from the AAP.

The educational characteristics of entrants into the AAP are assumed to be determined by current and past values of the relative educated wage. The relative decision wage (w_d) is defined as a weighted average of past educated wage relatives, the weights reflecting the relative importance of particular past educated wage relatives in the determination of the educational attainments of current new entrants into the AAP. Thus, at time t the decision wage is

$$w_{dt} = \sum_{i=1}^{8} b_i w_{t-8+i} \tag{5}$$

where

$$\sum_{i=1}^{8} b_i = 1.$$

Since the cohort of entrants in period t is composed of all persons who are 24 years old in t (who will be fully established members of the AAP in $t + 1$), the decision wage is a composite of relative wages over the age range 17–24 (i.e., $t - 7$ to t). In the absence of any conceivable method of estimation the set of specific weights as seen in table 4 is assumed. The very substantial weights attached to ages 18–20 reflect the

---

[12] In an earlier unpublished paper in which this model of educational adaptation was first developed, several alternative characterizations of technological change consistent with observations of the period 1956–72 are developed. Relying on logistic, exponential, and linear patterns of secular change in the parameters of a Cobb-Douglas technology, these provide a range of alternative projections of the future educational attainments of the AAP. The CES case, however, provides a useful focus of discussion: (*a*) the pattern of technological change which it generates lies between the extremes of exponential and logistically decaying development, (*b*) the higher elasticity of substitution (3.8 CES vs. 1 Cobb-Douglas) does not exhibit the extreme swings in the relative educated wage which are observed in the other cases, while (*c*) the broad pattern of educational adaptation observed in the CES simulation is qualitatively similar to that projected on the basis of the alternative assumptions.

TABLE 4

AGE-SPECIFIC WEIGHTS FOR RELATIVE DECISION WAGE (w_d)

i	Age	Wage Weight (b_i)	Weight Relative to Average
1	17	0.09375	0.75
2	18	0.1875	1.5
3	19	0.25	2.0
4	20	0.1875	1.5
5	21	0.125	1.0
6	22	0.0625	0.5
7	23	0.0625	0.5
8	24	0.03125	0.25
Total	...	1.0	...

NOTE.—$w_{dt} = \sum_{i=1}^{8} b_i w_{t-8+i}$.

importance of the initial decision to attend and persist in college, while the marginal weights for the period 22–24 reflect the fact that at this stage the bulk of the educational investment is a sunk cost.[13]

Given the decision wage, it is necessary to specify the way in which this wage variable is translated into educational attainments. Rather than assuming that the educated proportion of new entrants is itself a function of the decision wage, the model assumes that the change in that proportion is a function of the decision wage relative to some critical value, w', such that if $w_{dt} = w'$ the educated proportion of current entrants will be the same as the educated proportion of the immediately preceding cohort of entrants. Thus, w' can be interpreted as an equilibrium wage relative, that wage relative which is consistent with an unchanging educated proportion of new entrants.[14] The most appealing characteristic of this specification is that it permits the achievement and perpetuation of any educated share of the AAP at an ultimate wage relative w'.

Thus, s_t, the proportion of new entrants which is educated in time t, is viewed as a function of w_d, w', and s_{t-1}, of the general form

$$s_t = f(w_d;\, w',\, s_{t-1}), \tag{6}$$

with the characteristics (as indicated in fig. 1): $s_t = s_{t-1}$ if $w_{dt} = w'$;

---

[13] This specification of the decision wage clearly assumes static expectations, i.e., that entrants (and "preentrants") assume at each point in time that the current structure of relative wages by age will persist unchanged into the future, regardless of the degree to which this structure has changed in the past. No attempt has been made to incorporate more rational expectations.

[14] If the cost of education is a function of relative wages, then this specification of the equilibrium wage relative is perfectly defensible. Since the major costs are foregone earnings and the wages of faculty (i.e., higher education is labor intensive, with in excess of 75 percent of costs accounted for by employee compensation), this assumption seems justified, although it ignores possible changes in public subsidy policies.

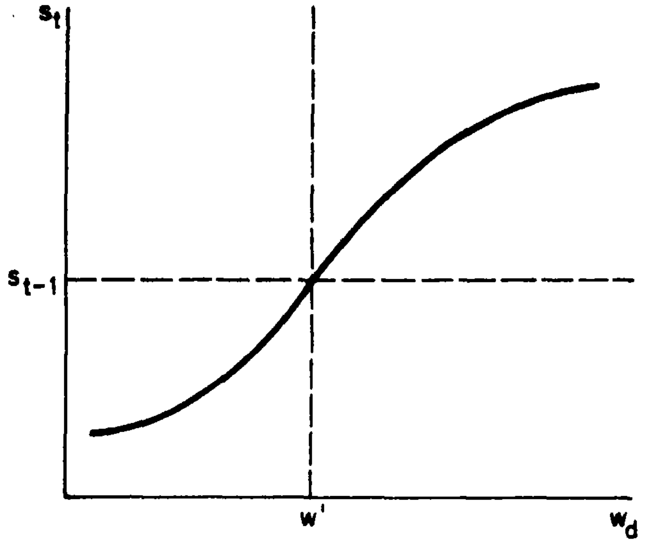

FIG. 1.—General relationship between relative educated wage (w) and educated proportion of entrants (s).

$\partial s/\partial w_d > 0$ (i.e., the educated share of entrants is a positive function of the decision wage); $\partial s/\partial w_d^2 \gtrless 0$ as $w_d \lesseqgtr w'$ (i.e., if the decision wage is less [greater] than the equilibrium wage, the educated share of entrants rises at an increasing [decreasing] rate with increases in the decision wage).

A number of specific functions can be devised which are consistent with these general specifications. For the declining proportion educated segment (decision wage less than its equilibrium level and hence $s_t < s_{t-1}$) the form utilized in the simulations is

$$s_t = ve^{qw_d} \text{ for } w_d < w', \tag{6a}$$

where v represents a lower limit on the educated proportion of entrants (i.e., the proportion of the entering cohorts which would seek education and enter the educated AAP even if the decision wage were zero). In this expression q is obtained by substituting s_{t-1} for s_t and w' for w_d:

$$s_{t-1} = ve^{qw'},$$
$$q = [\ln (s_{t-1}/v)]/w'.$$

In the case of an increasing educated proportion of entrants ($w_d > w'$ and hence $s_t > s_{t-1}$) the adjustment function is assumed to take the specific form

$$s_t = u\{1 - \exp [-q(1 + \ln w_d)]\} \qquad \text{for } w_d > w', \tag{6b}$$

where u represents an upper limit on the educated proportion of entrants (i.e., the maximum proportion of a cohort which is mentally and physically capable of education), and q is again obtained by substituting s_{t-1} and w' for s_t and w_d:

$$s_{t-1} = u\{1 - \exp[-q(1 + \ln w')]\},$$
$$q = -\{\ln[(u - s_{t-1})/u]\}/[1 + \ln w'].$$

It can be easily verified that these specific functional forms have the desired properties.

The effect of the previous educated proportion (s_{t-1}) on the relationship between s_t and w_d is of particular interest. In the case of the downside adjustment (eq. [6a], $w_d < w'$),

$$\frac{\partial^2 s_t}{\partial s_{t-1}\partial w_d} = \left[\left(\frac{vw'}{s_{t-1}}\right)\left(\frac{s_{t-1}}{v}\right)^{w_d w'}\right] + \left[\left(\frac{vw_d^2 w'^2}{s_{t-1}}\right)\left(\frac{s_{t-1}}{v}\right)^{w_d w'}\left(\ln\frac{s_{t-1}}{v}\right)\right] > 0$$

(i.e., the educated share is more responsive to the decision wage the higher the previous educated proportion).

In the case of upside adjustment (eq. [6b], $w_d > w'$), the responsiveness of s_t to any change in the decision wage is more complexly (and interestingly) related to the lagged value of the educated proportion, s_{t-1}:

$$\frac{\partial^2 s_t}{\partial s_{t-1}\partial w_d} = \frac{1}{w_d u(1 + \ln w')}\left(\frac{u - s_{t-1}}{u}\right)^{(1 + \ln w_d)/(1 + \ln w')} \times \left[1 + \ln\left(\frac{u - s_{t-1}}{u}\right)\right].$$

The first two terms of this expression are clearly positive. The third term, however, will be positive for very low values of s_{t-1} but negative for high values. Specifically, for values of s_{t-1} less than the critical value s^*, defined by $s^* = (1 - e^{-1})u \simeq 0.632u$, increasing the prior proportion educated increases the wage responsiveness (the more the system educates, the easier it is to increase the proportion educated in response to an increased educated wage relative), while beyond that critical value increasing the prior proportion educated reduces wage responsiveness. Interpretively, when the educational system is very elitist, educating only a fraction of new entrants, wages operate very ineffectively to attract additional entrants. However, as the system initially expands toward one of mass education, it becomes progressively easier to attract

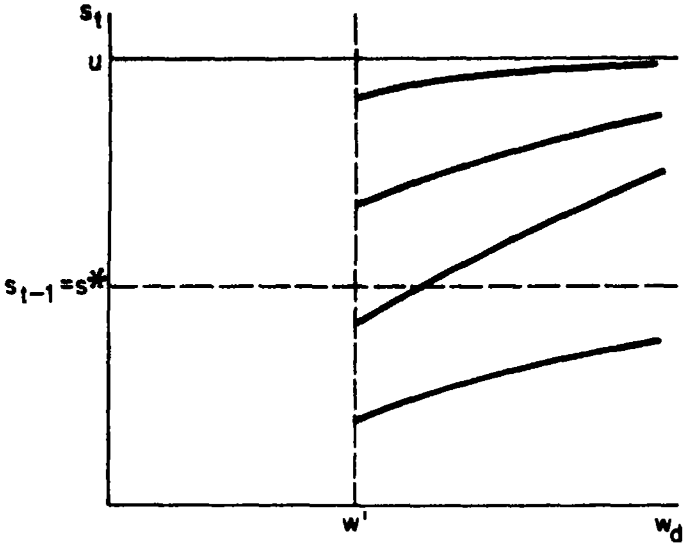

FIG. 2.—Wage responsiveness of increasing educated share of entrants dependent upon previous educated share.

additional entrants. But eventually the system approaches saturation, having attracted virtually all potential entrants capable of completing educational programs, and beyond this point wage responsiveness declines. This pattern of wage responsiveness is depicted graphically in figure 2.

Note that the form of the upside function requires that the critical value s^* occur at slightly less than two-thirds ($1 - e^{-1} \simeq 0.632$) of the upper limit regardless of the value of this limit. Thus, until the proportion of entrants educated reaches 63 percent of the maximum proportion, wage responsiveness increases. But as the system attempts to move beyond the 63 percent level, wage responsiveness declines. This pattern of response, if not the precise point at which the shift occurs, contains at least a basic plausibility.

To implement equations (5a) and (5b), it is assumed that the lower limit on the educated proportion of any cohort (v) is 0.05 and that the upper limit (u) is 0.65 (implying that upside wage responsiveness will begin to decline after 41 percent of entering cohorts have been educated). The equilibrium wage, w', has been determined so as to replicate experience over the period 1960–72, with a resultant value of 1.45.

Educational Attainments of Exits from the AAP

At the present stage this is the simplest component of the model. Because the focus is on population rather than labor force, it is not necessary to take into account changing patterns of labor-force participation. The only way a member of the AAP can exit is through death or retirement. In each model period the age distribution of the educated and uneducated components of the AAP are known. In moving to the next period all 64-year-olds are assumed to die or retire. Each younger age group is reduced in accord with age-specific mortality rates and aged 1 year. Previous 25-year-olds are replaced by the current cohort of entrants, adjusted for the effects of mortality between the ages of 24 and 25. The model currently does not employ differential mortality rates related to educational attainment, but this improvement could be easily introduced.

III. Simulation and Quasi Projection: 1960–2005

One component of the simulation, the underlying demographic context, does not represent a fabrication, and an understanding of this context is a necessary prerequisite to the evaluation of the simulation. [15]

The model projects the educational characteristics of the active adult population through the year 2005. All cohorts entering the AAP through the year 1997 (age 24 when their educational attainments are determined) have in fact already been born, and alternative projections of births for the period 1974–81 (entering cohorts of 1998–2005) differ very little from one another. While deaths must be projected, it is not unreasonable to anticipate that adult age-specific death rates will not change significantly over the next 35 years. Thus, the basic demographic context is subject to very little prediction error.

The most significant demographic characteristic of the period 1960–2005 is the succession of (*a*) very low birth rates of the 1930s, resulting in a gross flow into the AAP of only slightly more than 2 million persons per year prior to 1965; (*b*) the acceleration in births during the years of World War II, increasing the gross inflow to over 2.5 million per year in the late 1960s; (*c*) the explosion of births following the war, leading to a gross annual inflow of about 3.5 million during the early 1970s, increasing slowly to over 4 million in the early 1980s; and (*d*) the decline in births after 1960, resulting in a decline in the size of entering cohorts to about 3.5 million in 1990, with a gradual return to 4 million in 2000.

In contrast to these radical swings in births, exits (deaths and retirements) rise gradually from 1.9 million in 1960 to 2.6 million in the late

---

[15] The data underlying the simulation consisted of the following: (1) a vector of the proportion of people educated by age (25–64) derived from U.S. Bureau of the Census 1964, pp. 1–405; (2) a vector of population (i.e., the number of people at any given age) derived from ibid., pp. 1–349; (3) a vector of mortality rates for 1969, derived from U.S. Bureau of the Census 1973, p. 58; (4) a vector of entrants (24-year-old population) for the years 1960–2009 derived from Series E projections, U.S. Bureau of the Census, Series P-25, pp. 40–42, and U.S. Bureau of the Census 1964, pp. 1–349; (5) a 7-year lagged vector (1953–59) of the ratios of expected lifetime income by educational attainment, assuming a 3 percent discount rate and 2 percent annual increase in productivity, as derived from the U.S. Bureau of Census, Series P-60, pp. 38–39, 46–47, 50–51, 54–55, 58–59, 62–63, 66–67, and 70–71.

TABLE 5

CES PROJECTIONS OF EDUCATIONAL COMPOSITION AND MEAN AGES

	PERCENT EDUCATED				MEAN AGE	
YEAR	Actual AAP (1)	Equilibrium AAP (2)	Projected AAP (3)	Projected Entrants (4)	Educated AAP (5)	Uneducated APP (6)
1960...	9.1	12.7	9.1	12.8	40.4	43.2
1965...	10.4	14.0	10.2	17.8	40.9	43.8
1970...	12.1	15.4	12.0	23.5	39.9	44.0
1975...	...	16.9	15.0	29.2	38.1	43.6
1980...	...	18.5	18.7	32.9	36.9	43.0
1985...	...	20.2	22.5	33.1	36.8	42.5
1990...	...	22.1	25.0	26.2	37.7	42.2
1995...	...	24.1	25.9	19.3	39.7	42.1
2000...	...	26.2	25.7	16.0	42.1	42.1
2005...	...	28.4	24.8	15.7	44.4	42.2

1980s, decline to 2.3 million in the late 1990s (again due to the depressed birth rates of the 1930s), and then rise gradually to 2.5 million in 2005.

One of the consequences of the uneven rates of entrance and exit is a cyclical movement in the age structure of the AAP. The low rate of gross inflow in the early period led to an increase in the average age of the 25–64 population from 42.9 in 1960 to 43.6 in 1967. Thereafter the average age of the AAP declines, reaching its 1960 level in 1974 and continuing downward to a trough of 41.0 in 1988, after which it again begins to increase. By 2005 it has reached 42.8. As will be observed later, the average age of the educated AAP also follows this type of cyclical pattern, only in a more pronounced fashion.

Educational Attainment Simulation

Since the driving force of the system is the change in the production function generating the relative educated wage, it is useful to initiate an examination of the simulation with an overview of the technological dynamics. These can be most easily summarized by examining the secular changes in the equilibrium educated proportion of the population (i.e., the proportion which would result in an educated wage relative just equal to its equilibrium value). This is indicated in column 2 of table 5.

The CES technology, although exhibiting a limiting value of 100 percent of the AAP educated, generates a relatively gentle growth path of the equilibrium educated share of the AAP (s_t^e):

$$s_t^e = \frac{1}{1 + [(\gamma/\beta)e^{(g-h)\rho t}/w']^{-1/(1+\rho)}}.$$

From a 1960 value of 13 percent, this technology justifies an increase in the educated share of the AAP to 28 percent in 2005.

The columns 3 and 4 of table 5 present the results of combining the technological dynamics underlying column 2 with (*a*) the demographic realities of the period 1960–2005 and (*b*) the behavioral stipulations of the model. Columns 3 and 4 identify the educational characteristics of the entire AAP and of entering cohorts, respectively. For comparison, column 1 indicates the actually observed educated share of the AAP in 1960, 1965, and 1970. Because significant interest attaches to the changing age distribution of the educated and uneducated adult populations, columns 5 and 6 summarize the pattern of change in the mean age of the AAP by educational attainment.

The dynamics of the process of educational adaptation can be most easily comprehended by reference to table 6, which focuses on the patterns of movement in significant variables over the entire period. Over the course of the 1960s the actual proportion of the AAP educated fell significantly short of its equilibrium level (e.g., 10.2 percent actual versus 14 percent equilibrium in 1965). As a result of the consequent excess of the educated wage relative (1.59 in 1965) over its equilibrium value of 1.45, the share of entrants educated virtually doubled from 12.8 percent in 1960 to 23.5 percent in 1970; correspondingly, the number of educated 24-year-old entrants increased from 300,000 to 700,000.

After 1970, however, the gap between the actual and equilibrium educated proportion of the AAP rapidly narrows, as a result of, first, the rapid increase in the size of entering cohorts, reflecting the postwar baby boom, and second, the accelerator nature of the increase in the proportion of entrants educated. The latter stems from the combination of several factors: (1) Current entrants are responding not to the current wage relative but to the higher wage relatives prevailing 4–6 years in the past, as reflected in the weighted average relative decision wage. (2) Even though the wage relative is declining, as long as it continues above its equilibrium level the educated share of successive cohorts continues to rise. And (3) the increasing educated share of entrants has not yet passed the critical value $s^*[= (1 - e^{-1})u]$ above which wage responsiveness declines, so that the decline in the relative decision wage is at least partially offset by the increase in wage responsiveness which follows from increases in the previous educated share of entrants.

By 1979–80 the actual educational composition of the AAP reaches alignment with its then current equilibrium composition, with between 18 and 19 percent of the AAP educated (see fig. 3). At this point the relative educated wage has fallen to its equilibrium level. However, the educated share of entrants is significantly above this level, at in excess of 30 percent, and vastly above the educated share of exits from the AAP (less than 10 percent educated). Furthermore, the educated share of

TABLE 6

CES PROJECTIONS: CRITICAL VARIABLES AND DATES

	PERCENT EDUCATED		NO. ENTRANTS EDUCATED (MILLIONS)	RELATIVE EDUCATED WAGE	RELATIVE DECISION WAGE	MEAN EDUCATED AGE (YEARS)	MEAN UNEDUCATED AGE (YEARS)
	AAP	Entrants					
Initial (1960)	1960/ 9.1	1960/12.8	1960/0.3	1960/1.60	1960/1.60	1960/40.4	1960/43.2
1970.	1970/12.0	1970/23.5	1970/0.7	1970/1.56	1970/1.59	1970/39.9	1970/44.0
Equilibrium	1979/18.0	. . .	. . .	1979/1.45	1984/1.45	. . .	. . .
Peak	1998/25.9	1983/33.6	1983/1.4	. . .	. . .	. . .	. . .
Trough	. . .	2003/15.5	2001/0.7	1989/1.39	1993/1.39	1984/36.7	1996/42.1
Equilibrium	1999/25.8	. . .	. . .	1999/1.45	2003/1.45	. . .	. . .
Terminal (2005) . . .	2005/24.8	2005/15.7	2005/0.7	2005/1.52	2005/1.47	2005/44.4	2005/42.2

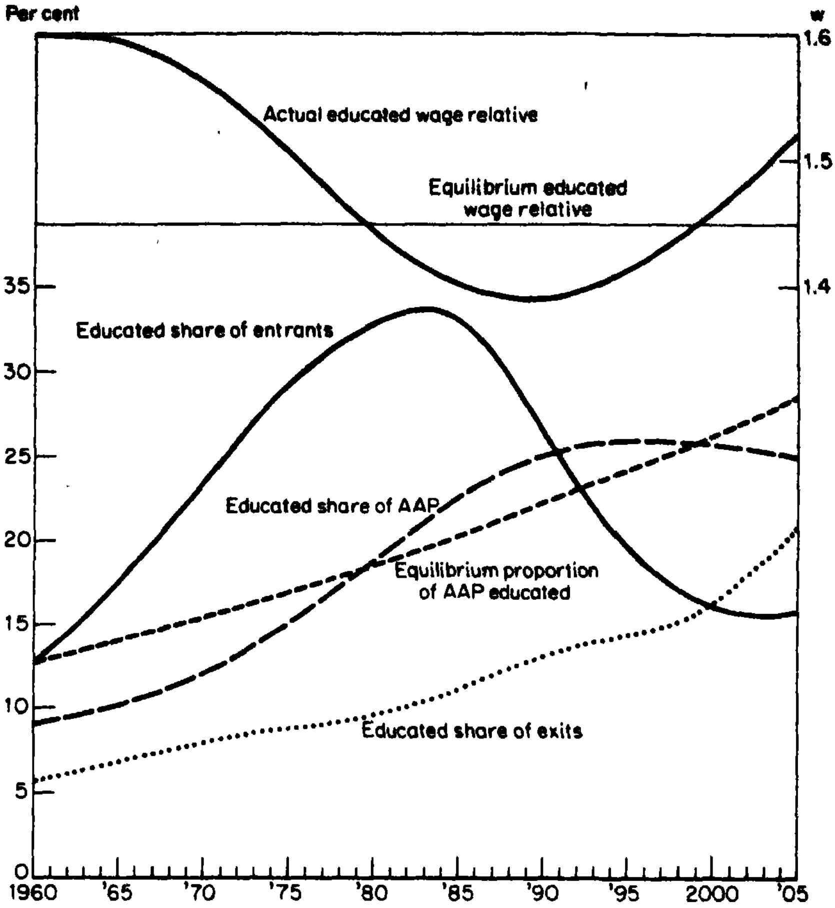

FIG. 3.—CES projections of educated shares of entrants, of exits, of AAP, and equilibrium proportion of AAP educated.

entrants continues to rise, peaking at 33.6 percent in 1983, due to the lagged decision wage, which follows the actual relative wage by about 4 years.

After 1983 the educated proportion of entrants begins to decline, but continues to surpass the educated share of the AAP, actual and equilibrium, until the early 1990s. The relative educated wage continues to decline to its trough of 1.39 in 1989 as the gap between the actual and equilibrium educated AAP widens. After 1989 this gap narrows as the educated share of entrants continues to decline, while the equilibrium educated share of the AAP rises toward the virtually stable actual educated share of the AAP.

The actual and equilibrium educational compositions of the AAP again come into alignment in 1999, with a relative wage equal to its equilibrium value. At this point, also, the educated share of exits for the first time exceeds that of entrants, leading to an absolute decline in the educated proportion of the AAP, while the equilibrium educated share continues to rise. Thus, the educated wage relative is again driven above its equilibrium level, and, with a 4-year lag, an upturn is induced in the educated share of entrants, which had bottomed out at 15.5 percent in 2003.

Thus, the process of educational adaptation over this 45-year span is one of cycling about a moving equilibrium, with the amplitude of this cycle determined primarily by the underlying wavelike movement in demography. That is, the lags in the educational adjustment process, which would generate a degree of disequilibrium in any event, are magnified by the succession of expanding and contracting entering cohorts. The expansion comes just as the size of entering cohorts is abnormally large, pushing the system toward vastly inflated educated attainments, while the contraction occurs when entering cohorts are themselves contracting rapidly, leading to an overreaction in the direction of reduced educational attainments. This contraction is itself reinforced by the earlier overexpansion, as mortality begins to claim significant fractions of the educationally inflated cohorts, driving up the educated share of exits from the AAP.

Projections of College Enrollment

Unfortunately, the model projects future educational attainments of 24-year-olds, not enrollment. However, the size of educated entrant cohorts in any period can serve roughly as a proxy for the size of college and university enrollments 3–6 years in the past, on which basis the quasi projections of table 1 were developed. Thus, the increase in educated entrants from 300,000 in 1960 to 700,000 in 1970 suggests an approximate 140 percent increase in college enrollments between the late 1950s and the late 1960s. This compares favorably with the observed increase in degree-credit enrollment of about 120 percent between 1956 and 1966. It would be expected that the change in degree attainments would overstate the change in enrollment over this period since the same factors driving up enrollment rates would also serve to reduce attrition from degree programs. This may in fact constitute one of the elements explaining initial increases in wage responsiveness with increases in the prior proportion educated: not only does college become available to a greater proportion of the cohort; in addition, the exclusivity of college programs is reduced, permitting a greater proportion of entrants to survive the educational gauntlet. Of course, the most direct consequence of persistently high monetary payoffs to higher education would be a reduction in voluntary attrition rates, as reflected in the positive weights attached to postentrance wages in generating the relative decision wage.

The peaking in the size of educated entrant cohorts in 1983 at 1.4 million suggests a doubling in enrollments between roughly 1966 and 1979, with the bulk of the increase taking place prior to 1975 and virtual stability between 1975 and 1980. Thus, the period of expansion would appear at present (1974) to be virtually at an end. Over the decade 1970–80 the model suggests a cumulative increase in degree-credit enrollment at about 40 percent, an increase which is roughly consistent with the NCES and Carnegie projections, ranging from 31 to 54 percent. However, after 1980 the model-based projections are clearly much more radical, with a projected 40 percent contraction between 1980 and 1990, compared with the Carnegie Projection II contraction of 8 percent. And between 1990 and 2000 the model projection of a further 12 percent decline contrasts with the Projection II increase of 26 percent.

Thus, over the full period 1970–2000 the model of educational adaptation suggests a contraction of 33 percent, while the Carnegie Commission anticipated an overall increase in enrollment of between 50 and 90 percent. In effect, while the declining size of entering cohorts is somewhat offset by the assumption of increasing enrollment rates in the *t-d* analyses, in the adaptation model the decline in the size of the college-age population is reinforced by the decline in the rate of college completion, from a high of 34 percent in 1983 to a low of 15.5 percent in 2003. Thus, the number of educated entrants declines from its peak of 1.4 million to about 700,000 in 2001. This suggests that the level of enrollment in the late 1990s will be about one-half its 1980 peak, and approximately equal to its level in the late 1960s. Thus, in effect, the entire net addition to the capacity of the higher-education system over the decade roughly 1967–77, equal to the capacity in 1967, according to the CES-based projections will be redundant by the year 2000.

It should be noted that the differences between the Carnegie and adaptations-model projections may be partially explained by the inclusion in the former and exclusion in the latter of nondegree-credit enrollment, a component which is expected by the commission to become an increasingly important part of total enrollment over the period. Also, the apparent change in attrition rates which led the model to overpredict the change in enrollment during the highly expansionary period 1956–66 may similarly lead it to underpredict enrollment growth (overpredict enrollment declines) during the post-1980 period of contraction. That is, declining incentives to the acquisition of a college degree will operate to increase attrition as well as reduce initial enrollment, and the ratio of enrollment to degree awards may significantly increase.

In any event, the next several decades will constitute a significantly traumatic period for traditional higher education. Even stability in enrollment would pose serious difficulties; a contraction, whether by 8 percent or by 46 percent, will necessitate much more severe readjustments.

IV. Social Concomitants of the Process of Educational Adaptation

In conclusion, it is useful to highlight briefly several of the more important social consequences and concomitants of the simulated process of educational adaptation.

A major dimension within which the consequences of this process give rise to concern involves the changing age distribution of the population, as depicted in figure 4. Over the first half of the 1960s the average age of both the educated and the uneducated AAP increased due to the low birth rates of the 1930s and the decline in college enrollments in the early 1950s, immediately following the postwar GI Bill surge. The average age of the uneducated AAP continues to rise until about 1970, but after 1965 the age of the educated AAP declines precipitously until the early 1980s. From a peak of 40.9 years in 1965 the mean educated age declines to a trough of 36.7 years in 1984. Note that this 4+-year decline is in excess of 10 percent of the 25–64 age range of the AAP. In the meantime the age of the uneducated AAP declines very gently from a 1971 peak of 44 years to 42.6 in 1984. Thus, the spread in average age between the educated and uneducated widens from 3 years in 1965 to 6 years in 1980. The potential social instabilities which might be implied by this relative youthening of the educated AAP, should, however, be mitigated by the decline in the relative educated wage, which falls below its equilibrium (replacement) value in 1979 and remains depressed until the year 2000. Thus, the relative saturation of the educated-labor market (tightening of the uneducated-labor market) should reduce the tensions which would otherwise arise as an older, uneducated populace was subjected to a progressively younger, educated population.

The conjunction of a declining educated age and a saturation of the educated labor market may result not only in the avoidance of social tensions between the educated and uneducated but may also carry positive benefits in its own right. Declining age and increasing competition within the educated-labor market may have marked consequences for the vitality of intellectual activity. For the higher-education sector, however, the experiences of this period may be quite the reverse. Given such institutional constraints as tenure arrangements, a serious process of atrophication may occur. Even on the basis of the Carnegie Commission's relatively optimistic enrollment projections, and assuming "no changes from present practices in retirement policies, student-staff ratios, net

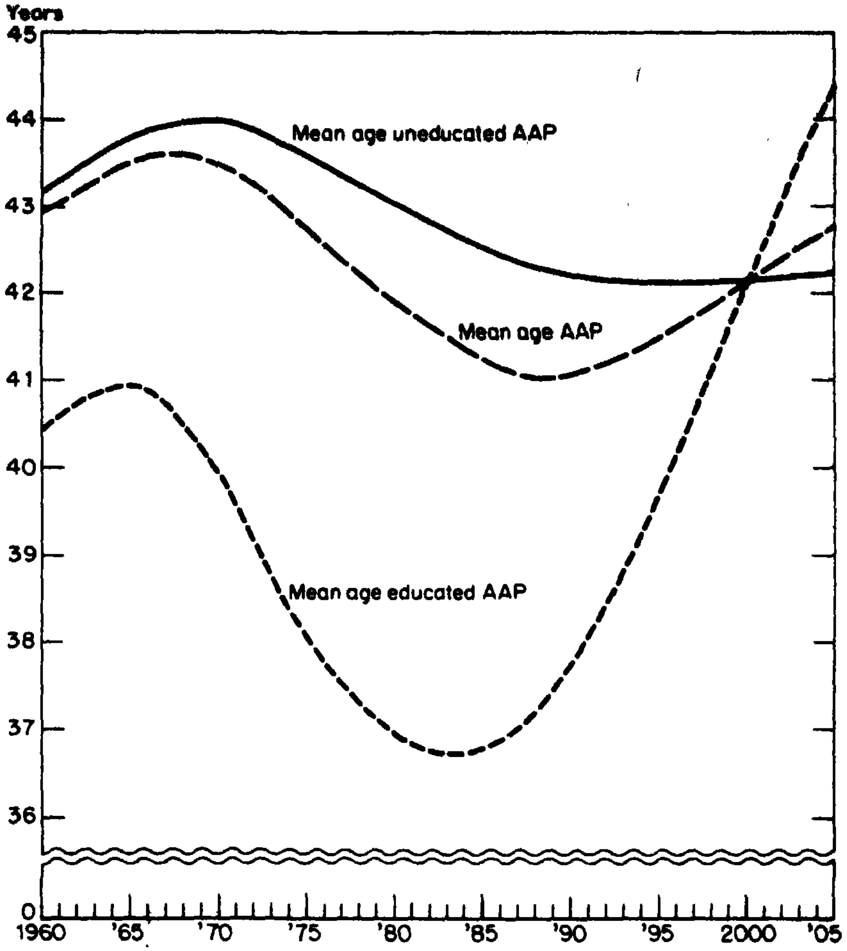

Fig. 4.—CES projections of the mean age of the educated and uneducated active adult population.

flows to employment outside academic institutions, and the like," the commission projects an increase in the proportion of faculty over age 50 from 23 percent in 1980 to 53 percent in 2000. Similarly, the proportion under 35 is projected to decline from 28 percent in 1970 to 4 percent in 1990.[16] The consequences of these potential developments for graduate education and for the research functions of the university may well be traumatic.[17]

---

[16] Carnegie Commission (1973, table B-10, p. 119).

[17] These potential consequences of enrollment contraction for higher education, and especially for the research university, are explored in detail in Dresch (1974).

After 1985, however, the general situation is radically altered. The mean age of the educated adult population increases steadily, reaching 44.4 years in 2005 (an 8-year increase from the previous trough, equal to 20 percent of the 25–64 range), while the mean age of the uneducated AAP declines to 42.1. And over this period declines are registered first in the educated proportion of entering cohorts and then in the educated proportion of the entire AAP. Thus, by 2005 an ever younger uneducated population is subjected to an ever older educated population. As one consequence of these developments, the intellectual dynamism of the preceding period may be followed by a period of intellectual stagnation as an aging educated population is insulated by a tight educated-labor market from any incentives for dynamically productive activity. Again, however, the condition of the higher-education sector may differ significantly, especially after 2000 when the inflated 1960s generation of young faculty is progressively eliminated by retirement and death, while enrollment finally begins to rise in response to the emergence of a significant excess demand for educated labor. Thus, the stage may well be set for a second "Golden Age," rivaling the euphoric period of the 1960s, in which graduate enrollments grew at record rates and, at least partly in consequence, the universities' share of research activity increased dramatically.

The full implications of the projected change in the age composition of the educated and uneducated populations, however, cannot be comprehended apart from the pervasive evolutionary consequences of the aging society. While the average age of the AAP (aged 25–64)will begin to rise after the mid-1980s, as the cohorts of the postwar baby boom pass 40 and are succeeded by the contracting cohorts of the 1960s, the most severe impact of postwar demographic trends will not be felt until after the year 2000, when the most swollen cohorts enter their 60s.

Table 7 examines probable changes in the relative size and composition of the dependent population over the period 1970–2050. "Dependent" in this context obviously has restricted meaning, referring to those of very young or relatively advanced ages, in general those not in the labor force. Thus, significant changes in patterns of retirement or of adolescent emancipation could greatly alter the real meaning of any age composition.

In general, the period 1970–2010 is *one of declining* relative dependency, with the young and old together declining in importance from 44 percent of the population to 37 percent. The composition of the dependent population, however, undergoes major change: Aged dependency is roughly constant while overall dependency declines, due to a decline in the proportion under 18 from 34 to 26 percent (1970–2010).

The under-18 population continues a slow, relative decline after 2010,

TABLE 7

"DEPENDENT" PROPORTION OF U.S. POPULATION, 1970–2050
(%)

Year	Under 18 and 65 or over	Under 18	65 or over	62 or over
1970	43.9	34	9.8	12.3
1980	40.6	30.2	10.4	12.9
1990	40.8	29.8	10.9	13.4
2000	38.3	27.6	10.6	12.7
2010	36.9	26.2	10.6	13.6
2020	38.7	25.6	13.1	16.5
2030	40.5	24.8	15.7	18.6
2040	39.7	24.8	14.9	17.9
2050	40.0	24.9	15.1	18.4

NOTE.—"Dependency" assumes continuation of current emancipation and retirement patterns. Estimates based on Census Series E population projections.

but just as the average age of the 25–64 population rises after the mid-1980s, the aged proportion of the population begins a precipitous increase after 2010. The proportion 65 or over actually increases by almost 50 percent between 2010 and 2030, rising from 10.6 to 15.7 percent. Because the cohorts reaching 65 over this period are those which constituted the most highly educated entrants into the AAP during the 1980s, the combination of increasing size of the retiring cohorts relative to the size of entering cohorts, with a deficit of educated labor, will create severe labor-market strains if the educational adjustment process functions as it has in the past.

More generally, without basic changes either in the welfare (relative incomes) of the retired or in their relative self-maintenance (labor-force participation), the rapid rise in the size of the aged population will impose an increasingly heavy burden on the economically active population, implying potentially serious social instabilities. However, the excess demand for educated labor which can be anticipated over this period may provide an effective mechanism for avoiding or mitigating these instabilities. Specifically, a rising relative educated wage, and a deficit of educated manpower, which cannot be easily eradicated by increasing the educational attainments of relatively contracted entering cohorts, may encourage highly educated potential retirees to delay retirement and may also result in general institutional changes favorable to later retirement, for example, elimination of mandatory retirement ages, reduction of work disincentives in social security and other retirement programs, etc. In any event, the implications of the historical process of educational adjustment can only be understood in the context of these broader social developments.

References

Bowles, S. "Aggregation of Labor Inputs in the Economics of Growth and Planning: Experiments with a Two-Level CES Function." *J.P.E.* 78 (January/February 1970): 68–81.

Carnegie Commission on the Future of Higher Education. *Priorities for Action.* New York: McGraw-Hill, 1973.

Cartter, A. "A New Look at the Supply of College Teachers." *Educ. Rec.* 46 (Summer 1965): 267–77.

Crean, J. "Foregone Earnings and the Demand for Education: Some Empirical Evidence." *Canadian J. Econ.* 6 (February 1973): 23–42.

Denison, E. "The Shift to Services and the Rate of Productivity Change." *Survey Current Bus.* 53 (October 1973): 1–37.

Dougherty, C. "Estimates of Labor Aggregation Functions." *J.P.E.* 80 (November/December 1972): 1101–19.

Dresch, S. P. "The College, the University and the State: A Critical Examination of Criteria for Institutional Support in the Context of Historical Development." Paper presented at a Conference on "The Role of the Undergraduate College in the Development of a Human Future," Emory University, July 1–2, 1974.

Fallon, P. R., and Layard, P. R. G. "Capital-Skill Complementarity, Income Distribution and Output Accounting." *J.P.E.* 83 (April 1975): 279–302.

Griliches, Z. "Capital-Skill Complementarity." *Rev. Econ. Statis.* 51 (November 1969): 465–68.

———. "Notes on the Role of Education in Production Functions and Growth Accounting." In *Education, Income and Human Capital*, edited by W. Lee Hansen. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1970.

National Center for Educational Statistics (NCES), U.S. Department of Health, Education, and Welfare. *Projections of Educational Statistics to 1980–81.* Washingtion: Government Printing Office, 1972.

———. *Projections of Educational Statistics to 1981–82.* Washington: Government Printing Office, 1973.

Parnes, H. *Forecasting Educational Needs for Economic and Social Development.* Paris: OECD, 1962.

Psacharopoulos, G., and Hinchliffe, K. "Further Evidence on the Elasticity of Substitution between Different Types of Educated Labor." *J.P.E.* 80 (July/August 1972): 786–92.

Tinbergen, J. "Substitution of Graduate by Other Labor." *Kyklos* 27, no. 2 (1974): 217–26.

U.S. Bureau of Census. *Current Population Reports.* Series P-20, nos. 45, 77, 99, 121, 138, 158, 169, 172, and 182. "Educational Attainment: (March 1952, 1956, 1958, 1962, 1964, 1965, 1966, 1967, and 1968)." Washington: Government Printing Office, 1953, 1957, 1959, 1963, 1965, 1966, 1967, 1968, and 1969.

———. *Current Population Reports.* Series P-25, no. 470. "Projections of the Population of the United States by Age and Sex: 1970 to 2020." Washington: Government Printing Office, 1971.

———. *Current Population Reports.* Series P-60, no. 74. "Annual Mean Income, Lifetime Income, and Educational Attainment of Men in the United States for Selected Years, 1956 to 1968." Washington: Government Printing Office, 1970.

———. *Statistical Abstract of the United States 1973.* Washington: Government Printing Office, 1973.

———. *U.S. Census of Population: 1960.* Vol. 1, *Characteristics of the Population.* Pt. 1, United States Summary, sec. D, "Detailed Characteristics." Washington: Government Printing Office, 1964.

———. *U.S. Census of Population: 1970.* Vol. 2, *Subject Reports.* Pt. 7B, "Industrial Characteristics." Washington: Government Printing Office, 1973.

Welch, F. "Education in Production." *J.P.E.* 78 (January/February 1970): 35–59.

The Production and Distribution of Cognitive Skills within Schools

---

Byron W. Brown and Daniel H. Saks

Michigan State University

Do school inputs matter? Regression analysis of Michigan Assessment Survey Program data (1971) shows inputs are productive, but only under our new criteria. Many previous economic studies have concluded that school inputs do not matter because mean school output is often uncorrelated with input variations. But if every student's performance matters to a school, then productive inputs affect the entire distribution of student performances. If that distribution can be characterized by its mean and variance, then inputs are productive if *either* moment is affected. One cannot, however, tell how much these distributional effects depend on technology and how much on tastes of school authorities.

Introduction

There is a burgeoning literature applying economic concepts to the analysis of schools. Typically, a school is thought to be a firm whose output is the mean score of its students on some achievement test. Most of these studies use regression techniques to estimate the productivity of various school inputs, and most have concluded that school inputs do not matter. It is a peculiar doctrine to be embraced by so many educators. In this paper, we have tried to rethink the microeconomics of the problem. We consider the problem in which each student's performance matters to the school, which, as a multiproduct firm, must allocate its resources among students in some rational way. In such a model, mean student

This work was supported by the Center for Urban Affairs, Michigan State University, and the National Bureau of Economic Research. We particularly wish to thank our MSU colleague Wilbur Brookover for both intellectual and financial assistance and Robert Huyser of Michigan's Department of Education for provision of the data. Also, we would like to thank John Kain, John Meyer, and others in the Harvard Seminar on Urban and Regional Studies; Jan Kmenta of the University of Michigan; Anita Summers of the Federal Reserve Bank of Philadelphia; and an anonymous referee for sophisticated and helpful comments on an earlier draft.

[*Journal of Political Economy*, 1975, vol. 83, no. 3]

© 1975 by The University of Chicago. All rights reserved.

performance can be unaffected by variation in inputs, even when those inputs are productive. Thus earlier regression results may well be an inappropriate test of input productivity. While the proper subject of the theory is the entire distribution of outputs or students' performances, distinctions must be made between the distributional consequences of technology and those of tastes or values. It is important to examine whether there are biases in the technology of education or in the preferences of school decision makers leading to choices of a wider dispersion of student outcomes (elitist systems) or to a narrower dispersion (leveling systems). These are difficult matters to disentangle with available data and techniques. We present some empirical estimates which are generally consistent with our theory and which show that, for our data, most inputs do indeed matter in most places. Had we neglected the distributional questions raised by our theory, we would have come to the very opposite conclusion.

Theoretical Model

The economic problem in the educational process can be formulated very simply: How should the school authorities use their resources to improve certain attributes of their students? Our concern here is to explore some of the logical economic consequences of education being a production process applied to *individual* students in classes in schools. It is not so simple to jump from this basic disaggregate formulation to a relationship (so often estimated) in which the mean score on some standardized test of students in a school depends on certain school characteristics.[1] The average output of the educational process should not be considered without simultaneously analyzing the distribution of that output within the school.

A General Optimization Model

It will be useful (in part because it is so rarely done) to set out a very general economic model of resource allocation in a school. This more comprehensive formal model provides the framework for the more specific discussion to follow. As economists, we imagine a school's choices to be the result of a conflict between values and opportunities.[2] Schools are complex nonprofit organizations in which teachers, administrators, parents, the political authority, and students all make or participate to varying degrees in decision making and in which historical forces impinge upon current choices. We only require that the complex objectives or values of such an organization be summarized by a well-behaved welfare function (w) in which the attributes (S_j for the jth student) of each of the N students in the school appear with positive first derivative. These attributes are measured as scores on some standardized test of cognitive skills.[3] Although much has been written about how schools operate, there seems to be little or no systematic analysis of how individual student performances get weighted by different kinds of schools, and it will turn out to be an important question. Postulating a welfare function is our way of saying that schools think of themselves as being primarily in the business of teaching the things they test.

---

[1] For a review of this literature, see Averch et al. (1972).

[2] Of the basic approaches surveyed by Averch et al. (1972), the "organizational approach" corresponds most closely to a study of the school's values or objective function, and the "process approach" corresponds to the opportunities or the production constraint. The essence of our theory is that the third approach cited by them, the "input-output approach," is really the result of the working out of the strains between the first two approaches. We will refer to the "input-output approach" as the reduced form.

The school's opportunities are determined by production and resource-allocation constraints. There is a production relation for each student (f_j for the jth student) in which the student's score is determined by how much of each of the M different kinds of inputs (various types of teachers, physical equipment, and other facilities) get applied to the particular student (I_{kj} being the amount of input k applied to student j). The production function will also include the initial endowment of skills (E_j) already possessed by the student. These endowments could be thought of as an index of the quality of the student's entering skills and depend, in turn, on a variety of environmental and other factors.[4] Since a student's learning is affected by the quality of fellow students, the endowments of the other students in the school or class should be included in the production function for each student. Given our primary interest in the way inputs get used within a school, we assume that the amount of each of the M inputs is already fixed at the beginning of the school year.[5] There is an input-allocation constraint (g_k) for each of the M inputs which describes the possible ways in which inputs can be distributed across the students. The simplest form of such a constraint would be where one unit of input k devoted to any student j reduces the amount of input k available to the other students by one unit (i.e., the total amount of an input is the simple sum of the amounts used on each student, so the partial derivative of g_k with respect to I_{kj} is $+1$). However, inputs are often applied so that one student's receipt of an input unit will not reduce the inputs available to other students by an equal amount (e.g., students are usually taught in classes), and so we leave the function (g_k) in a general form.[6]

---

[3] The expression S_j is more likely to be a vector of scores measuring attributes as diverse as knowledge of the multiplication tables and fervor in saluting the flag. A radical critique of American education (see Gintis 1971) has been that schools are really allocating resources in order to accomplish "proper" socialization of the student rather than greatest possible cognitive development. Because of the data available to us, we concentrate on the development of cognitive skills as measured by a standard test and represented as a scalar. One hopes its availability is an indicator of its importance as an educational objective. The existence of other outputs could create problems for our analysis insofar as some of the same inputs which develop cognitive skills also perform socialization functions. In the extreme, the argument would be that if schools were rational and if they placed no value on cognitive development, then inputs would not be allocated to optimize education in cognitive skills and our model would be irrelevant to observed behavior. That, in principle, is an empirical question. The absence of an observed effect of inputs on mean cognitive achievement scores would not, as developed below, be a confirmation of the radical position because multiple outputs (e.g., mean, variance, kurtosis) are involved even when schools only produce cognitive skills. If cognitive skills along with other student attributes are valued by schools, then our analysis will be biased only to the extent that either the efficiency of the inputs in producing the different kinds of outputs or the relative valuation by school boards of those diverse outputs vary systematically across school districts. In other words, there has to be a systematic substitution in either the production or consumption of these outputs which cannot be statistically controlled before we need be concerned about a partial analysis of the outputs of the educational system. In the absence of knowledge on this question, we are forced to ignore it.

To summarize, the school tries to maximize some welfare function of student scores subject to N production constraints and M input-allocation constraints. The school must decide how to allocate its inputs in order to achieve the desired *distribution* of student scores or outputs. The optimization problem can be written directly as one in which the following function is maximized:

$$\begin{aligned} L = {} & w(S_1, \ldots, S_N) \\ & - \sum_{j=1}^{N} \lambda_j [f_j(I_{1j}, \ldots, I_{Mj}; E_1, \ldots, E_N) - S_j] \qquad (1) \\ & - \sum_{k=1}^{M} \alpha_k [g_k(I_{k1}, \ldots, I_{kN}) - I_k], \end{aligned}$$

where λ_j and α_k are Lagrange multipliers.

---

[4] We are sweeping another problem under the carpet. Simply because schools specialize in education does not mean that a very large part of a student's education takes place there. The home, the streets, and the workplace may still provide the bulk of the education in our society. We are here concerned only with difference in skills caused by the school part of the educational system. But if there is systematic, uncontrolled variation across school districts in the way the nonschool parts of the educational system perform, then a partial analysis of the school sector of the educational system will be misleading.

[5] For an example of an analysis in which the amounts of the inputs are endogenously determined, see Brown (1972).

[6] The introduction of the g_k is a way of recognizing the existence of joint products (but not necessarily in fixed proportions). It is, as well, a specification of the mode of production in schools in that tutorial and lecture systems of instruction would have different forms of the g_k's. The ready existence of the special linear form of g_k (i.e., no constant term and weights all equal to one) assures us that no sensible school would ever teach in classes unless such teaching were more efficient, so these constraints are certain to be convex.

Assuming there is an interior maximum,[7] the first-order conditions tell us that schools will allocate the different inputs to the different students so that marginal value (or social) products adjusted for opportunity costs are equal for all students and all inputs.[8] In other words, the school's decision regarding the distribution of outputs in a school will depend upon three variables: the marginal products of the available inputs when applied to particular students, the marginal utility of particular students' performances as evaluated by the school authorities, and the amount of any input exhausted when that input is applied to a particular student under the existing conditions of production.[9] These first-order conditions can be manipulated to find reduced-form equations for the S_j which will depend on endowments and productive conditions. The essential point is that the schools are making decisions about the whole distribution of outputs or test scores of which the mean is but one parameter.

The school, then, is like a petroleum refinery which takes crude oils of different characteristics (e.g., low sulfur, high sulfur, etc.) and produces a mix of final products ranging from tar to high-octane gasolines. The mix of final products depends on the relative prices of different outputs, the productivity of various processes applied to particular types of crude oil, and the prices or opportunity costs of various inputs. To the extent that technologies get embodied in capital, the refineries will display short-run variations in productivity as the mix of crude oil types changes.[10] Hence, there is an important dynamic component to this assignment problem. Since teachers' skills, physical plants, textbooks, institutions, and equipment are in the short run fixed, this dynamic problem affects

---

[7] As we show below when fig. 3 is discussed, corner solutions are plausible, and restrictions on the curvature of these functions are, of course, required.

[8] These marginal conditions can be stated in a variety of equivalent ways. It is easy to give an economic interpretation to the following:

$$\frac{(f_{\sim j})_{\sim k \sim j}/(g_{\sim k})_{\sim k \sim j}}{(f_j)_{\sim kj}/(g_{\sim k})_{\sim kj}} = \frac{(f_{\sim j})_{k \sim j}/(g_k)_{k \sim j}}{(f_j)_{kj}/(g_k)_{kj}} \qquad \text{(all } j, k\text{)}, \tag{a}$$

$$\frac{\partial w/\partial S_j}{\partial w/\partial S_{\sim j}} = \frac{(f_{\sim j})_{k \sim j}/(g_k)_{k \sim j}}{(f_j)_{kj}/(g_k)_{kj}} \qquad \text{(all } j, k\text{)}, \tag{b}$$

where $(.)_{kj}$ denotes the partial derivative of a function with respect to changes in the amount of the input k applied to student j. The notations $\sim k$ and $\sim j$ refer, respectively, to any input other than k and any student other than j. Condition (a) means that any input should be allocated to each student to the point where the marginal product of that input (adjusted for the opportunity costs of applying the input to that student rather than another) is the same as the marginal product of any other input when applied to that student. Condition (b) means that inputs should be allocated to students so that the ratio of the marginal products of those inputs when applied to any two students (again, adjusted for costs) be in inverse proportion to the additional welfare generated by a unit increase in the scores of those students.

[9] This is the place where interest in student discipline or socialization would enter into the considerations even of a school board which is only interested in developing cognitive skills.

[10] For a discussion of the petroleum problem, see Manne (1956).

schools too. It may, for example, take some time for a school to adapt technology to teach students from new and different subcultures. If schools are particularly rigid, they may never adapt.

So far, our analysis tells us that there is a reduced-form equation for each student, in which output or the student's score depends upon the distribution of endowments and the specific applications of inputs (i.e., we can use the optimality conditions to solve for output in terms of inputs and endowments). Unfortunately, we do not have data on the entire distribution of outputs for particular school districts, and even if we had, we would have to aggregate to the school or district level because it is almost impossible to measure accurately the quantity of an input applied to a *particular* student (i.e., we can measure I_k but not I_{kj}).[11] Furthermore, the points which we wish to make here can be brought out without going beyond a discussion of the most basic parameters of the output distribution. To make any further progress, we are going to have to make some assumptions about the form of the distribution and of the functions appearing in the model.

The Welfare Function

Taking our cue from portfolio theory, we will assume that the welfare function can be approximated by a function containing only the mean and variance of the achievement scores for all of the students. Thus a school's decisions will, for our purposes, be described in terms of choices about the mean and variance of achievements of their students. One common rationale in portfolio theory for this type of welfare function would be individual student scores entering the social welfare function in quadratic form.[12] Unlike the usual formulation of the portfolio selection

---

[11] The problem is that these inputs get applied to different kinds of students by a variety of mechanisms ranging from formal "tracking" to the sort of informal methods described so well by Kozol (1967). These procedures are embedded in our production function.

[12] For example, if we select

$$W = w(S_1, \dots, S_N) = \sum_{j=1}^{N} (aS_j + bS_j^2),$$

then

$$\frac{W}{N} = a\bar{S} + \frac{b}{N} \sum_{j=1}^{N} S_j^2.$$

Since

$$\text{var}\ (S) = \frac{1}{N} \sum_{j=1}^{N} S_j^2 - \bar{S}^2,$$

we can substitute to get $W/N = a\bar{S} + b[\text{var}\ (S) + \bar{S}^2]$. Thus $W/N = w[\bar{S}, \text{var}\ (S)]$ under these assumptions. The reader should be aware that we are making a virtue of necessity here. We only have data on the mean and variance of the student scores for our school districts. If the reader does not like the quadratic utility justification for the mean-variance analysis which follows, he might at least grant us the fact that a weighted average of normal variates will itself be normally distributed and can be fully described

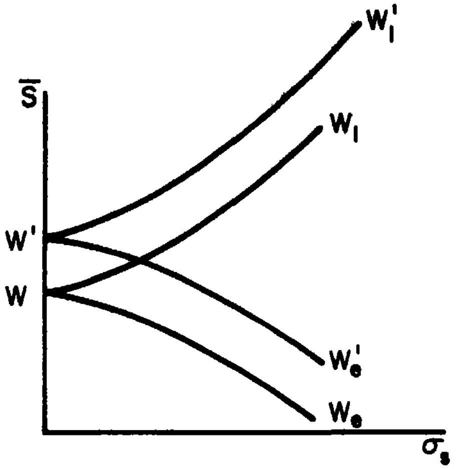

FIG. 1.—Welfare contours for leveling and elitist school boards

problem, the welfare indifference curves of school authorities in mean-standard deviation space (where σ_s is the standard deviation of student scores) are no more likely, a priori, to be concave upward than downward. In figure 1, the upward-sloping curves represent the preferences of a school board which, in selecting between two policies producing the same mean achievement scores for its students but different dispersions, prefers the one with a smaller dispersion. Call such school boards *levelers* (risk averters in the parlance of portfolio theory) and label their indifference curves WW_l. (The "prime" indicates the curve with the higher level of welfare.) The downward-sloping curves (WW_e) represent the preferences of school boards which, *ceteris paribus*, prefer wider distributions of achievement. Let us call these boards *elitists* (risk lovers in portfolio analysis). The indifference curves could be straight horizontal lines, which would imply a school board that pays no attention to the dispersion of its students' achievements.[13] We have no reason to think any of these possibilities is unlikely.

---

by the mean and variance. Since the tests given to students are usually designed to provide a normal distribution of outcomes for the state, we may be on safe ground with this assumption. Even so, individual schools may value skewness, and this ought to be investigated. Portfolio theory is described in a number of places, but an article by Tobin (1965) is most helpful because he is careful to include the preferences of "risk lovers" in his discussion, and this is not an unlikely case for our purposes. For a critique of the mean-variance approach, see Samuelson (1967).

[13] Since the slope of the indifference curves is $-2b\sigma_s/(a + 2b\bar{S})$ for the case of a quadratic utility function and since we have assumed that welfare increases with the scores of individual students over any relevant range, we can characterize school boards in terms of the sign of b. For levelers, $b < 0$; for elitists, $b > 0$. If $b = 0$, the welfare function is not quadratic and variance is not a factor in the school board's decision.

What does it mean to say that school authorities value either increases or decreases in the variance of achievement of their students? There have been two major and potentially conflicting goals of American public education. On the one hand, there is the democratic ideal of the old common school stressing equality of educational attainment. On the other hand, there is the efficiency goal of developing or enhancing individual differences among students to fit them for a differentiated economy and society or, as President Eliot of Harvard once put it, "to sort the pupils and sort them by their evident or probable destinies."[14] What should be obvious about these goals is that they imply propositions about the variance of achievement rather than the mean of achievement. If Einsteins or even bank presidents are valued disproportionately by school boards, then we will observe a wider dispersion of scores within a school. But if every student is expected to come up to some minimal level and if all students are exposed to the same instruction, the dispersion will be narrower. Our earlier analysis, however, warns us to be careful about making such statements because the optimal allocation requirements also depend on the marginal productivity and marginal cost of inputs applied to particular students. Even a school board which had a "leveling" set of goals could, if faced with the right technology and student body, still end up producing a wider dispersion of outputs if that were the bias of the current technology. Thus we cannot easily separate the production biases from the consumption biases of a school system. This question of the biases of the technology in education is not a subject on which there is very solid information (in part because it is a distinction which seems most likely to occur to an economist). There is some evidence that current technology embodied in the training of teachers, classroom materials, and notions of acceptable teaching practices in schools may well be biased toward student differentiation irrespective of the desires of the school authorities.[15] Consequently, we will have to spend some time developing the production-possibilities part of our problem.

The Production Set

The production constraint now must specify the limits of mean and variance which can be realized in a school and depends not only on the inputs available but also on the initial educational endowments of the students and the distribution of those endowments. Let us begin with a

---

[14] Quoted in a paper by some of our colleagues at the Center for Urban Affairs who, coming from the field of educational sociology, have reached many conclusions consistent with our work. In their paper, they set out the rationale of the two ideal types of schools and review the evidence from the literature in sociology, social psychology, and education (see Brookover et al. 1974). The work of Bowles (1971) is also relevant here.

[15] Brookover et al. (1974) take much the same perspective and provide some evidence on the technology question.

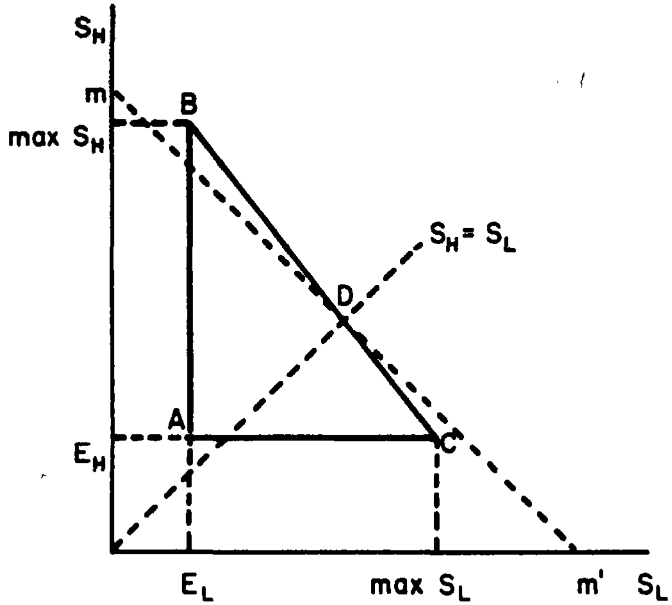

FIG. 2.—A production-possibilities set for two students

simple example. Suppose we have a school with just two students (or homogeneous groups of students) who differ in their cognitive achievement levels at the beginning of the school year and also in their productivity responses to a given application of inputs (i.e., an input has a different marginal product for each of the students). Let the production relations for the two students be linear:

$$\begin{aligned} S_H &= E_H + \alpha_H I_H \\ S_L &= E_L + \alpha_L I_L, \end{aligned} \tag{2}$$

where S is the achievement score produced by the application of some quantity of input I to a student with an initial endowment or score of E. The marginal product of I for that student is α. The subscript H designates the student who begins the year with a high achievement score, and the subscript L designates the student with the low initial score. Total inputs available, $\bar{I}$, equals $I_H + I_L$.

The set of production possibilities is shown in figure 2 as the triangle ABC. Since higher scores are valued for all students, the school will choose from points on the line segment BC unless resources are wasted. If all of the school's resources are applied to student H, the result is point B. If everything is devoted to student L, point C is the result. As drawn, the absolute value of the slope of BC is greater than one implying that I is more productive when applied to H than L (i.e., $\alpha_H > \alpha_L$). Our problem is mapping the given production-possibilities set from student achievement space (S^N) into mean–standard deviation space where we want to depict the choice.

In order to help our intuition about how to do this transfer, we draw two additional lines in figure 2. The 45° line from the origin represents all points where the scores of the two students will be equal. This is also the locus of points where the standard deviation of the students' scores will be zero. We also draw a line mm' perpendicular to the 45° line and going through the intersection of the 45° line and the production-possibilities frontier. Since the slope of this line is -1, it represents the locus of points corresponding to a constant mean. There is, naturally, a family of parallel lines or isomean curves, and those lying farther from the origin represent a higher mean of student scores. We can now see how to draw line *BC* in mean–standard deviation space. As we move from point *C* to *D*, the mean is increasing and the standard deviation decreases to zero. From *D* to *B*, the mean continues to increase, as does the standard deviation. Figure 3*a* shows the production-possibilities curve of figure 2 in mean–standard deviation space. It is piecewise linear (in mean-variance space it would be a parabola). When we draw indifference curves on the same graph, we can see that only the segment of the production frontier from *D* to *B* would be relevant for the given problem.[16] If a school board were elitist, all resources would be devoted to the high-scoring student. If it were composed of levelers, then resources would be allocated to produce both the lower mean and standard deviation represented by point *f*. In the case where the marginal product of an input is greater for the student with the lower level of initial endowments, the curve *BC* will be flipped about the point *D* as shown in figure 3*b*. Now, the segment *DC* is the only relevant part of the production frontier except for the special case of figure 3*b*, where initial endowments are sufficiently maldistributed and the school board is sufficiently elitist to still choose to devote all resources to the student starting off with a higher score.[17] More generally, we can see

---

[16] In the particular example, this is equivalent to saying that a school board will not act so as to change the relative rankings of students during the course of the year. Notice that it would be possible for initial endowments to be so unequally distributed at the beginning of the year and productivity to be such that the 45° line never intersects the triangle *ABC*. In that case, *BC* would be a straight line in mean–standard deviation space which does not touch the ordinate.

[17] The fact that we need only consider the part of the frontier between the 45° line and the point where all resources get used on the student for whom inputs are most productive is a result of the peculiar symmetry of the quadratic utility function about the 45° line in student achievement space. The slope of the indifference curve in student achievement space (i.e., the rate of commodity substitution) is $dS_H/dS_L = -(a + 2bS_L)/(a + 2bS_H)$. The first-order conditions for an optimum tell us that the indifference curve will be tangent to the production-possibilities frontier at the constrained extremum. Since the slope of the production-possibilities frontier is minus the inverse ratio of the marginal products, we have at the solution $\alpha_H/\alpha_L = (a + 2bS_H)/(a + 2bS_L)$. Thus, the ratio of the outputs will be greater or less than one if and only if the ratio of the marginal products is, respectively, *greater or less than one.*

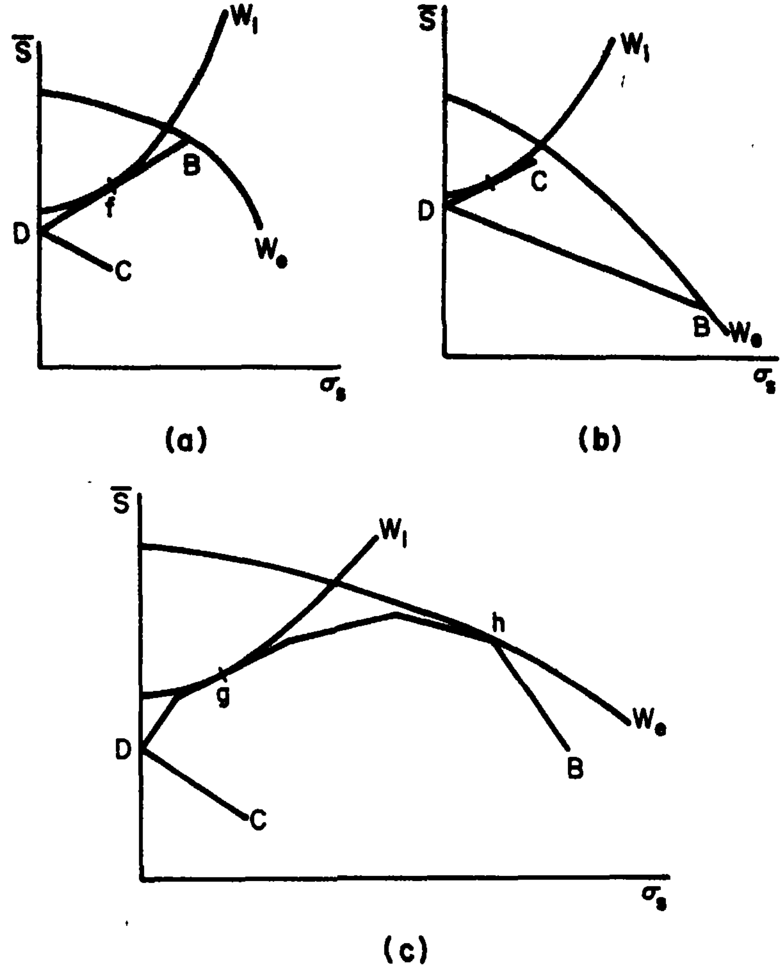

Fig. 3.—Equilibrium choices in mean–standard deviation space

that the greater the disparity in the relative marginal productivity of inputs applied to the two students, the steeper the slope of the production-possibility frontier in mean–standard deviation space. In the case of figure 3*a*, either developments in technology or purchases of inputs which are more productive when applied to the "poorer" students would cause a reduction in the slope of the production possibilities frontier and, *ceteris paribus*, a reduction in the mean and the standard deviation.

There are two important extensions of the model that must be mentioned. First, the number of inputs may be increased. This will produce a production-possibilities frontier in student achievement space which is concave to the origin and is similar to the production set in a linear programming problem.[18] The second extension of the model is to increase the number of students. Unfortunately, this produces a situation in which we no longer have a one-to-one mapping between the two spaces. The production-possibilities frontier now becomes an area in mean–standard deviation space. The utility function does at least guarantee that only points on the boundary of the area need be considered.

The Reduced Form

An important empirical result of many "production function" studies we alluded to earlier was that teachers and other inputs do not seem to matter for achievement. Under reasonable assumptions, we can now show that even if inputs are productive when applied to students (i.e., teachers matter and marginal products are positive), the mean level of achievement observed may not depend on inputs at all. In other words, a regression of mean achievement levels on input levels need not tell us anything at all about the productivity of inputs. To demonstrate this result, and others, we can specify a version of our model which has a special case where school inputs affect achievement in the true production function, yet mean achievement is independent of school inputs in the reduced-form equation for mean achievement. This version of the model will also serve as the basis for empirical estimation.

If an aggregate linear production function is assumed as an approximation of the production-possibilities frontier, then the required reduced form can be derived from an aggregate welfare function which is fully quadratic in both mean and standard deviation. We are still restricting ourselves to the first two moments of the output distribution. The school's allocation problem for N students and M inputs is now seen as the following:

$$\begin{aligned} \max w &= \sum_{j=1}^{N} S_j[a + (b - d)\bar{S}] + \sum_{j=1}^{N} S_j^2\left[d + \frac{c\sigma_S}{\sigma_S^2 + \bar{S}^2}\right] \\ &= a\bar{S} + b\bar{S}^2 + c\sigma_S + d\sigma_S^2 \qquad (3) \\ &\text{subject to } \alpha\bar{S} + \beta\sigma_S = \gamma_0 + \sum_{i=1}^{M} \gamma_i I_i . \end{aligned}$$

This is a helpful specification because the weights for S_j and S_j^2, the terms in brackets, depend on the mean and standard deviation of achievement; and it yields a rather neat form of the welfare function in $\bar{S}$ and σ_S. This produces two linear reduced-form equations for the mean and standard deviation:

---

[18] Such a frontier, mapped into mean–standard deviation space, is shown in fig. 3*c*. The important thing to notice is that the existence of inputs with differing comparative advantages will reduce the possibility that a school will specialize in better students no matter how elitist the school board. Beyond some point, the standard deviation can only be increased by lowering the mean, a case which is excluded from the one-input case.

$$\begin{aligned} \bar{S} &= K_S + \frac{d\alpha}{b\beta^2 + d\alpha^2} \sum_{j=1}^{M} \gamma_i I_i; \\ \sigma_S &= K_\sigma + \frac{b\beta}{b\beta^2 + d\alpha^2} \sum_{i=1}^{M} \gamma_i I_i, \end{aligned} \tag{4}$$

where K_S and K_σ are constant terms.[19]

One particularly important possibility can be demonstrated directly from the above formulation. If the marginal social welfare of changes in the dispersion of student achievements is constant, then $d = 0$, and mean achievement levels will be independent of even the most productive inputs. This is clear from the above equation where, if $d = 0$, the inputs drop out of the reduced-form equation for $\bar{S}$. In this world, one would expect a zero coefficient for inputs in a regression of mean achievement level on inputs.

There are a number of other things to note about this formulation. We defined "elitist" and "leveler" schools in terms of the curvature of the school's preference contours. Now, that curvature may change as the achievement distribution changes. The formal conditions are

$$-\left(\frac{c + 2d\sigma_S}{a + 2b\bar{S}}\right) \begin{array}{l} > 0 \text{ for levelers} \\ < 0 \text{ for elitists,} \end{array} \tag{5}$$

where the left-hand side of the inequality is the slope of the welfare function in mean–standard deviation space. Thus various kinds of allocation scenarios could be described in terms of the parameters a, b, c, and d. If schools, for example, like to put resources into remedial education first and then, as the worst students improve, put resources into enrichment programs, that will be consistent with a regime (among others) where $d < 0$ and other parameters are positive (and $2d\sigma_S$ eventually becomes less than c as the remedial programs reduce σ_S). This kind of switching is a useful property for the welfare function to have.

A second property of this formulation is that the effect of an input on the two outputs is complicated. If we observe only the reduced form by performing linear regressions of σ_S and $\bar{S}$ on I_i, then γ_i is never identified and we can only make some crude inferences about it. We can distinguish

---

[19] These are, to put it mildly, a mess:

$$K_s = \left[\left(\frac{c}{\beta} - \frac{a}{\alpha}\right) - \frac{2d\gamma_0}{\beta^2}\right] \frac{\alpha\beta^2}{2(b\beta^2 + d\alpha^2)};$$

$$K_\sigma = \left[\left(\frac{c}{\beta} - \frac{a}{\alpha}\right) - \frac{2b\gamma_0}{\alpha^2}\right] \frac{\alpha^2\beta}{2(b\beta^2 + d\alpha^2)}.$$

several cases. (1) If the coefficient of I_t is nonzero in *both* of the reduced-form equations, then γ_t must be positive (we assume it cannot be negative). (2) If the coefficient of I_t is nonzero in *either* reduced-form equation, then γ_t must be positive and the input probably only appears to be unproductive in one of the equations because of tastes (i.e., either b or d is zero). (3) If the coefficient of I_t is zero in both equations, then either γ_t is zero or both d and βb are zero (α cannot be zero by assumption). But if d and βb are both zero, then this implies a corner solution to the optimization problem, a situation which may be inconsistent with our model but is certainly a real-world possibility and is illustrated by point B in figures 3*a* and *b*.[20] In summary, it is only by excluding the possibility of corner solutions that these types of reduced-form equations could ever imply that an input were unproductive. A nonzero coefficient in *either* equation does assure us that an input is productive.

Another property of this formulation is that there is a restriction across the equations. The ratio of the coefficients of an input in the two equations must be equal to that ratio for any other input. That ratio is $(d\alpha)/(b\beta)$ for all inputs. This property is a consequence of picking such a simple, that is, linear, production constraint. The restriction says that (1) the relative productivity of any input on high- and low-endowment students is the same for all inputs and (2) that relative productivity is constant. Condition 1 is our earlier marginal condition in the general model, but condition 2 is imposed by the specific production constraint used. We could in theory impose this restriction in our estimation procedure,[21] but we would really like to see whether the data generate this result. It will be a test of the adequacy of the linear production relation as an approximation over the observed ranges. We now turn to the empirical results.

Some Estimates of the Model

If our theory is correct, then we can add stochastic terms and try to estimate the pair of reduced-form equations discussed in the last section (equation set [4]). One problem is how to treat the level and distribution of student endowments. We will assume, for the time being, that these endowment variables affect the constant term in each equation (i.e., they determine the position of point A in figure 2) but not the coefficients of the inputs. Since these endowment variables are primarily for control purposes here, we will have to experiment with some alternative specifications of the endowments to provide that surface which best fits the data.

---

[20] This noninterior solution can be visualized in mean–standard deviation space. If d and b are both zero, then the indifference curves are straight lines. If d is zero, then the indifference curves are parabolas with parallel vertical directrices; and if β is zero, an input is equally productive on both tails of the distribution, making the production-possibilities curve a horizontal line (i.e., independent of standard deviation).

[21] It involves a difficult nonlinear restriction across equations.

The data for our estimates were collected and organized by the Michigan State Department of Education in their Michigan Educational Assessment Program for the 1970–71 academic year. Our observations are classified according to the three broad community types: (1) cities, (2) suburbs, and (3) rural and towns.[22] Although this is an unusually fine data base, it still limits severely the kind of econometric analysis we can currently perform. Research is now underway to extend these data, but for the present we are forced to report results based on some rather naive econometric and theoretical assumptions. We forewarn the reader of this and will try to point out particular problems as we go along. For example, our model is implicitly formulated in terms of flows of outputs and inputs. The measures of outputs used here are at best only proxies for those flows: $\bar{S}$ is the mean level of composite achievement for fourth-graders in the school district, and σ_S is the standard deviation of composite achievement for fourth-graders in the school district.[23] There are at least two potential problems with these variables that have to be assumed away. First, schools have to use their resources so that there is no systematic variation over grade levels in input and output flows, or else our results would vary with the grade level observed. Second, there should be no systematic mobility of students among school districts. In essence, students cannot benefit from schools they do not attend or when they are constantly adjusting to new schools. If, for example, it were found that poorer children change schools more or less frequently than other families, this could bias our estimates for the effect of economic status and school inputs. Since the Michigan Educational Assessment was conducted for both fourth- and seventh-graders, our selection of the former group reflects our hope that these assumptions are more plausible for students who have had less opportunity to move among districts and have only been through elementary school.

There is a similar problem with the input and those endowment variables which refer to the entire school system in the district and not just to the particular grade being observed. The input variables used here are (1) the ratio of pupils to teachers and other professional personnel; (2) the average number of years of experience of the instructional staff; (3) the percentage of teachers with master's degrees; and, to capture

---

[22] Only those districts with schools teaching kindergarten through twelfth grade were included. The community types were defined as follows: city: a community of over 10,000 people serving as an economic focal point of its environs; suburb: a community of any population size that has a city as its focal point; town and rural: cities of less than 10,000 people or communities of less than 2,500 people.

[23] The composite achievement score is the arithmetic mean of a student's score on a test of reading, mathematics, and mechanics of written English. These scores were scaled to have a mean of 50 and a standard deviation of 10.

TABLE 1

WEIGHTED* MEANS AND STANDARD DEVIATIONS OF VARIABLES BY COMMUNITY TYPE

Variable	City	Suburb	Town and Rural
Mean achievement level (index)	49.01 (3.46)	51.56 (2.55)	51.00 (2.16)
SD of achievement level (index)	8.92 (0.65)	8.37 (0.62)	8.27 (0.75)
Socioeconomic status (index)	48.52 (3.14)	51.73 (3.63)	49.57 (2.02)
SD of socioeconomic status (index)	10.26 (0.80)	8.87 (0.71)	8.93 (0.92)
% white students (fraction)	0.74 (0.24)	0.96 (0.12)	0.97 (0.05)
Students in district (thousands)	50.94 (94.78)	8.87 (7.77)	2.36 (2.24)
Average experience of instructional staff (years)	10.56 (1.48)	8.14 (2.12)	9.45 (2.32)
Students/teachers and professional staff	24.44 (1.90)	24.19 (2.35)	24.42 (2.41)
% of teachers with master's degrees	0.33 (0.07)	0.31 (0.11)	0.19 (0.09)
No. of observations	38	116	365

* The weights are those used in the regressions below and are the square root of the number of students in the district.

possible economies or diseconomies of scale, (4) the number of students in the school district. The measures of student endowments used are (1) the mean level of a socioeconomic status index for the fourth-graders in the district,[24] (2) the standard deviation of that socioeconomic status index, and (3) the percentage of white students in the school district. The reader is reminded that these endowment variables are meant to control for various environmental factors such as being in a ghetto, nutritional deficiencies, schooling of parents, and a variety of other nonmeasured influences on cognitive performance. Table 1 shows the calculated means and standard deviations by community type for the variables used in our study.

The reduced-form equations were estimated by weighted least squares, and the results are reported in table 2. Because we were dealing with grouped data from the individual school districts, we used weighted least squares in which the weights were the square root of the number of students

---

[24] The socioeconomic index is meant by the sociologists who designed it to be a measure of social class and is based on levels of consumption, education, and type of family structure, among other things.

TABLE 2

Weighted Least-Squares Regressions of Mean and Standard Deviation of Achievement on Input and Endowment Variables by Community Type

Input and Endowment Variables	Community Type and Dependent Variable (DEP)					
	City		Suburb		Town and Rural	
	DEP = $\bar{S}$	DEP = σ_s	DEP = $\bar{S}$	DEP = σ_s	DEP = $\bar{S}$	DEP = σ_s
Average experience of instructional staff	0.23 (1.31)	−0.11* (1.71)	0.23** (2.84)	−0.13** (3.50)	0.25** (6.0)	−0.080** (4.90)
Students/teachers and professional staff	−0.19 (1.13)	0.0083 (0.13)	−0.085 (1.56)	−0.047** (2.00)	−0.055 (1.37)	0.039** (2.40)
% of teachers with master's degree	3.32 (0.72)	−0.13 (0.07)	0.014 (0.009)	1.62** (2.32)	3.29** (3.0)	−0.40 (0.91)
SES (index)	0.44** (2.52)	1.02 (1.14)	0.53** (11.5)	1.42** (3.98)	0.45** (8.4)	0.20 (0.36)
(SES)^2	...	−0.0098 (1.10)	...	−0.014** (4.09)	...	−0.0020 (.36)
SD of SES (index)	−0.17 (0.43)	0.13 (0.77)	−0.12 (0.62)	0.098 (1.22)	−0.045 (0.42)	0.12** (2.73)
% white students (fraction)	6.47** (2.90)	6.95** (2.84)	4.73** (4.1)	0.19 (0.08)	5.03** (2.73)	5.13 (1.13)
(% white)^2	...	−6.11** (3.01)	...	−1.08 (0.58)	...	−4.08 (1.49)
Students in district (thousands)	−0.0029 (0.67)	−0.00087 (0.56)	−0.0074 (0.43)	0.0095 (1.28)	−0.12** (2.69)	0.054** (3.04)
Constant term	25.92** (2.13)	−19.3 (0.90)	21.04** (5.39)	−26.6** (2.83)	22.98** (7.39)	0.78 (0.06)
$\bar{R}^2$	.84	.40	.75	.24	.35	.16

Note.—Absolute t-values are in parentheses.
* Significant at .10 level.
** Significant at .05 level.

in the district.[25] Our least-squares procedure assumes that the independent variables are predetermined and that schools are simply allocating the resources they are given. The world is obviously not so simple, and, for example, when we "observe" teachers "affecting" students, we are also observing certain kinds of students attracting or being assigned to certain kinds of teachers. Finally, we are not even able to take advantage of the probable correlation of the error terms in our two equations to increase the efficiency of our estimates because both equations have basically the same independent variables.[26]

All regressions were calculated separately for each community type, and no pooled results are presented because we rejected at the .01 level the hypothesis that the sets of estimated coefficients were equal for all community types.[27] This partitioning reflects the possibility that different kinds of communities have both different welfare functions and different production functions. The variation in production functions may seem a bit odd in view of the *apparent* uniformity of teacher training, but as the petroleum-refinery analogy suggests, if schools are relatively inflexible institutions, changes in the student body may move schools away from the frontier of their production constraint. City schools have, of course, experienced quite dramatic changes in their students over the past 20 years. A proper test of this kind of hypothesis would require data generated over time.

Productivity of Measured Inputs

The first and most important question to ask is whether the measured inputs (experienced teachers, teachers with master's degrees, and just teachers) are productive. On the criterion of whether these inputs have a positive coefficient in the mean-achievement regressions, we might conclude that they are not productive. After all, only three of those nine coefficients are significant. But on the criterion developed in our theory section, we come to the opposite conclusion. By that rule, if an input has a significant coefficient in either equation, then that input must have a positive marginal product. In seven of the nine cases, the input is significantly productive. In fact, *every* measured input is significant at the .05 level in both the suburbs and in the town and rural areas. It is only in the cities that some inputs are not significantly productive, and we have already suggested a possible reason for this. Yet even here, we should be careful. First of all, we showed that a zero coefficient in both equations need not imply a zero marginal product unless we exclude corner solutions. Second, there is so much multicollinearity among the independent variables in the equations for the cities that the standard errors of the coefficients are greatly inflated.[28] Thus we are applying a conservative test of significance. In short, inputs matter when properly evaluated.

---

[25] Since the error will be a function of the size of the population tested, ordinary least squares would be inappropriate because the stochastic terms would be heteroskedastic (see, e.g., Kmenta 1971, pp. 322–36).

[26] For example, we cannot gain efficiency using Aitken's method because the sets of regressors are the same in the two equations (see Kmenta 1971, pp. 516–29).

[27] We used a generalization of the Chow test to compare the three equations.

To evaluate the relative impacts of these inputs in the different reduced-form equations, it would be helpful to standardize the input coefficients. Economists naturally think about elasticities when faced with this problem, but in this case such measures are inappropriate. For the kinds of variables included in our analysis, percentage changes are not always meaningful. We would really like to scale our variables by the amount of variation one might typically observe for them in the world. This is true not only for the indices whose measures are so arbitrary, but also for some of the input variables which have remarkably little variation across school districts. This line of argument suggests that beta coefficients or weights are the relevant standardized measure.[29] These are displayed for the input variables in table 3. Readers will draw their own conclusions, but it seems true that experienced teachers have particularly strong impacts everywhere. Since they both raise the mean and lower the standard deviation everywhere, their net effect is to improve the worst students. The same effect holds, but less powerfully, for teachers with master's degrees and for teacher/student ratios in town and rural areas and cities. Master's degrees and better teacher/student ratios increase the standard deviation in suburbs, but since the mean is relatively unaffected, it is hard to conclude that the better students are more improved unless the worst students are actually hurt (or make less than normal progress).

These net distributional impacts are the result of both technology and tastes, and, of course, we cannot (with this kind of regression technique) disentangle the two types of effects. This becomes clear when one realizes that the distributional impact of a unit change in an input will be

$$\frac{\partial \bar{S}}{\partial I_i} \div \frac{\partial \sigma_S}{\partial I_i} = \frac{d\alpha}{b\beta},$$

as explained earlier. (These ratios are displayed in table 4.) We cannot tell, for example, whether the apparent trade-off between mean and

---

[28] The condition number (i.e., the ratio of the largest to the smallest eigenvalue) for the intercorrelation matrix of independent variables in the cities equation is very much larger than for the other equations. For a discussion of the effect of multicollinearity on variances of the estimated coefficients, see Kmenta (1971, pp. 380–91).

[29] These are the coefficients in a regression in which all the variables have been divided by their observed standard errors in this particular sample (see Goldberger 1964, pp. 197–200). There are some problems with this, since the distributions of some of these variables are obviously not Gaussian (e.g., the percentage variables can only range between zero and one).

TABLE 3

BETA WEIGHTS FOR INPUT AND ENDOWMENT VARIABLES IN TABLE 2

INPUT AND ENDOWMENT VARIABLES	COMMUNITY TYPE AND DEPENDENT VARIABLE					
	City		Suburb		Town and Rural	
	DEP = $\bar{S}$	DEP = σ_s	DEP = $\bar{S}$	DEP = σ_s	DEP = $\bar{S}$	DEP = σ_s
Average experience of instructional staff	0.097	−0.25*	0.19**	−0.43**	0.26**	−0.25**
Students/teachers and professional staff	−0.110	0.024	−0.078	−0.18**	−0.06	0.13**
% of teachers with master's degree	0.069	−0.014	0.0006	0.30**	0.14**	−0.049
SES	0.4**	4.92	0.75**	8.34**	0.42**	0.55
$(SES)^2$	...	−4.67	...	−8.46**	...	−0.55
SD of SES	−0.040	0.16	−0.033	0.11	−0.019	0.14**
% white students	0.45**	2.55**	0.22**	0.036	0.13**	0.38
$(\% \text{ white})^2$	...	−2.94**	...	−0.29	...	−0.49

* Significant at the .10 level.
** Significant at the .05 level.

TABLE 4

RATIOS OF INPUT COEFFICIENTS IN THE REDUCED-FORM EQUATIONS BY COMMUNITY TYPE

RATIOS*	COMMUNITY TYPE		
	City	Suburb	Town and Rural
Average experience of instructional staff..	−2.09	−1.77	−3.12
Students/teachers and professional staff ..	−22.89	+1.81	−1.41
% of teachers with master's degree ...	−25.54	+0.01	−8.22

* Precisely, $(\partial \bar{S}/\partial I_i) + (\partial \sigma_S/\partial I_i)$ for the ith input.

standard deviation in the allocation of experienced teachers reflects their efficiency with the worst students (α/β) or the schools' preferences in allocating them to these students (d/b). An implication of the reduced form was that this ratio would be the same for all inputs in a community type. From table 4, it seems that this is not the case. Since the welfare function is quite general, we might conclude that α/β is what varies for the different inputs. In fact, there is rather little we can infer from this behavior except that our assumption of a linear production constraint is not adequate.[30] It illustrates just how meager the results of these types of regression studies must necessarily be. The important questions are not whether inputs are productive but rather on whom they are productive and by how much. For the answer to those questions, other types of data and analysis will be required.

Effects of Measured Endowments

Although the endowment variables are not the focus of this analysis, there are several things that are worth noting here about their effects. As in most previous studies, race and socioeconomic status were found to be important variables. Their beta weights were highly significant and in general somewhat larger than those for the inputs in the mean equations.

The effect of class and race on the standard deviation of achievement turns out to be strongly nonlinear and significant in the metropolitan areas. Class seems to be the better endowment variable in the suburbs, and race seems to work better in the cities. The estimated relationship between the standard deviation of achievement and socioeconomic status in suburban school districts is shown in figure 4. The function is almost

---

[30] The reader should be reminded that this finding is inconsistent with the specification assumed earlier for evaluating our coefficients. We do not believe that these earlier conclusions are sensitive to this particular form of misspecification, though we feel that this is an area requiring further work.

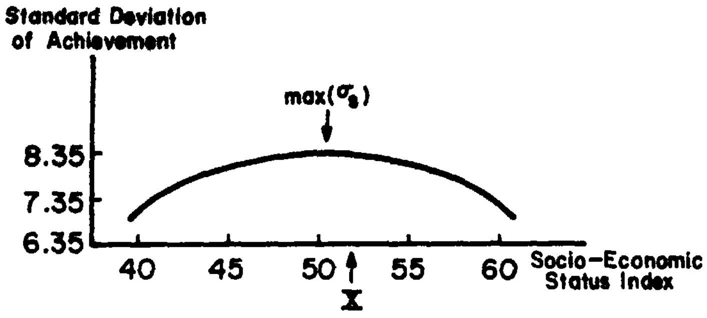

FIG. 4.—Estimated relationship between standard deviation of achievement and socioeconomic status index for suburban school districts (estimates based on coefficients reported in table 2 with all other variables evaluated at their weighted means).

symmetrical about the mean level of socioeconomic status. In fact, a linear specification of the relationship produces an insignificant coefficient for this index. A similar relationship with respect to race exists in the city schools. What this implies, if true, is that high- and low-class and segregated school districts tend to be more egalitarian than middle-class and integrated schools.

A remaining question to be asked of these endowment variables is whether they affect the coefficients of the input variables. To test for this, we added a group of multiplicative interaction terms in the inputs and endowments. According to an F-test at the .05 level, these did not add significantly to the explanatory power of the equations.[31]

Conclusions

We have argued here that there are good economic reasons for considering the entire distribution of student outputs in schools, and we have suggested that at least the standard deviation and the mean of that distribution should be included in an economic model of the school system. It is only in this multiple-output context that the productivity of measured school inputs can be evaluated. For example, previous researchers would have concluded that our measured inputs were unproductive, where we have tried to show that the opposite conclusion is warranted. But we have also tried to show that these kinds of regressions cannot in principle tell us anything about the size of the marginal products or an input's relative productivity with high- and low-endowment students. To answer those questions with nonexperimental data, it will be necessary to understand

[31] There were a few significant interactions. In the standard deviation equation for the suburbs, SES standard deviation had a positive effect with class size and percentage of master's degrees; and in the town and rural equations for the mean, SES standard deviation had a negative effect on average experience.

the nature of the schools' objective functions. It is only by understanding those objective or welfare functions in different kinds of schools that any progress can be made in unraveling the important empirical questions about the marginal productivity and distributional biases of inputs. In the generally available data, we observe only the interactions of technology and tastes.

References

Averch, Harvey A., et al. *How Effective Is Schooling? A Critical Review and Synthesis of Research Findings.* Santa Monica, Calif.: RAND, 1972.

Bowles, Samuel. "Unequal Education and the Reproduction of the Social Division of Labor." *Rev. Radical Polit. Econ.* 3, no. 4 (Fall-Winter 1971): 1–30.

Brookover, Wilbur B.; Gigliotti, R.; Henderson, R.; Niles, B.; and Schneider, J. "Equality, Differentiation, Testing, Assessment and Accountability in Education." In *Seventy-third Yearbook of the National Society for the Study of Education.* Chicago: Nat. Soc. Study Educ., 1974.

Brown, Byron W. "Achievement, Costs, and the Demand for Public Education." *Western Econ. J.* 10, no. 2 (June 1972): 198–219.

Gintis, Herbert. "Education, Technology, and the Characteristics of Worker Productivity." *A.E.R. Papers and Proc.* 61, no. 2 (May 1971): 266–79.

Goldberger, Arthur S. *Econometric Theory.* New York: Wiley, 1964.

Kmenta, Jan. *Elements of Econometrics.* New York: Macmillan, 1971.

Kozol, Jonathan. *Death at an Early Age.* Boston: Houghton Mifflin, 1967.

Manne, Alan S. *Scheduling of Petroleum Refinery Operations.* Cambridge, Mass.: Harvard Univ. Press, 1956.

Samuelson, Paul A. "General Proof that Diversification Pays." *J. Finance and Quantitative Analysis* 2, no. 1 (March 1967): 1–13.

Tobin, James. "The Theory of Portfolio Selection." In *The Theory of Interest Rates,* edited by F. H. Hahn and F. P. R. Brechling. London: Macmillan, 1965.

Nonmonetary Factors in the Demand for Medical Services: Some Empirical Evidence

---

Jan Paul Acton

RAND Corporation

Nonmonetary factors are expected to assume an increasingly important role in determining the demand for medical care as the out-of-pocket money price falls (due to spreading health insurance coverage or the enactment of the federal health insurance legislation). A utility maximization model is used to develop predictions for the demand for "free" and nonfree care. A simultaneous-equation system is estimated on a survey of users of New York City's "free" outpatient departments and municipal hospitals. The empirical results support the major predictions that nonmonetary factors such as travel distance will function as prices in discouraging demand and that earned and nonearned income have different impacts. A number of implications for public policy are suggested, including the possibility of substituting income maintenance for the direct provision or insurance of medical care.

I. Introduction

The demand for health care would deserve close scrutiny by researchers if only because of the size of the health sector. In addition to its size, the health sector has been one of the most inflationary in recent years, causing increasing interest by the public and policymakers alike. Finally, the growing numbers of proposals for significant changes in the health care system, especially proposals for federal health insurance legislation, make

This paper is a revised version of a RAND report, "Demand for Health Care When Time Prices Vary More than Money Prices," which was delivered to the winter meetings of the Econometric Society, 1972. The report was prepared for the Office of Economic Opportunity and the Health Services Administration of New York City. Raymond Lerner, of the State University of New York at Stony Brook, provided survey data for analysis. William Butz, David Chu, Michael Grossman, John Koehler, Joseph Newhouse, Charles Phelps, Norman Shapiro, Finis Welch, and a careful reviewer made comments on an earlier draft. Without trying to implicate them, I am grateful for their assistance.
[*Journal of Political Economy*, 1975, vol. 83, no. 3]

© 1975 by The University of Chicago. All rights reserved.

imperative a better understanding of the determinants of demand. This research is part of a growing literature that studies the demand for medical services using disaggregated data.[1] It differs from previous literature in concentrating on alternative mechanisms that may arise to determine demand for medical care as money prices to the patient fall through insurance.

The out-of-pocket money price of medical care has been decreasing as a proportion of total price in recent years, primarily because of the spread of health insurance and the rising opportunity cost of time (and perhaps increased time needed to receive care). There is every reason to believe that money prices will continue to fall in relative importance because of the secular trend in insurance coverage and opportunity cost of time and, perhaps even more important, the prospect of national health insurance. With the decreasing relative importance of money prices, it is reasonable to expect an alternative mechanism to control demand. A mechanism involving time is quite likely to assume this role since medical care usually requires a payment in both travel time and waiting time.[2] Indeed a substantial increase in queues and waiting lists was one of the widely noted results of the enactment of the National Health Service legislation in Britain (see Harris [1951] and citations therein).

This paper considers the effects of travel distance in determining the demand for medical services in New York City, an especially good "laboratory" in which to try to examine the effects of nonmonetary prices because of the long-standing availability of free care in the city's municipal hospitals and clinics.[3] After developing a formal model of the demand for medical services that includes a payment in money and in time for private care, the predictions are tested on a cross-sectional survey of about 2,600 users of city hospital outpatient departments (OPDs).

Although limitations of the data base indicate cautious interpretation, the empirical results lend support to the model's major predictions. Empirical verification of the conjecture that time is important in determining the demand for care raises a number of important policy issues. These not only include the effect on the distribution of services to recipients of care; they also indicate powerful policy instruments for increasing the medical access of target populations. Some of these policy options—which include the location of clinics and the substitution of income maintenance for subsidized care—are discussed at the end of this paper.

---

[1] See, among others, Grossman 1972*b*; Phelps and Newhouse 1972; Acton 1973*a*; and Rosett-Huang 1973.

[2] The importance of time as a determinant of demand was suggested by Becker (1965), its application to the demand for medical care by, among others, Leveson (1970) and Holtman (1972).

[3] In a related study, Acton (1973*a*) examined the role of travel time and waiting time in the demand for medical services using two surveys of poverty neighborhoods in New York. This study differs from the previous report in a number of ways. First, it is based on a survey of the users of outpatient departments (OPDs) rather than a survey of the general population in a neighborhood. Consequently, we know that everyone had nonzero utilization of the health sector. This allows the specification of a simultaneous set of equations to describe the demand at several different sources of care. Second, the chief measure of price paid for "free" public care is travel distance to the clinic. Further, travel distance is specified as endogenous in this system. In the previous study, the large proportion of observations on the limit value of zero health sector utilization forces the estimation of reduced-form equations using Tobit estimation technique.

II. Model of Demand for Medical Services

The underlying theoretical model for this study is developed in detail in Acton (1973*a*, 1973*b*).[4] The model concentrates on the role of money prices, time prices, and earned and nonearned income. For simplicity, the model is developed in terms of only one provider of services, but the implications for several providers can easily be drawn. In the theoretical and empirical sections to follow, I will be studying the demand for care at alternative providers—free city clinics, private physicians, and hospitals.

Assume two goods enter the individual's utility function: medical services, m, and a composite, X, for all other goods and services. Using an assumption of fixed proportions of money and time to consume m and X and the full income assumption, the model can be represented as follows:

Maximize

$$U = U(m, X) \tag{1a}$$

subject to

$$(p + wt)m + (q + ws)X \leq Y = y + wT, \tag{1b}$$

where

U = utility,
m = medical services,
X = all other goods and services,
p = out-of-pocket money price per unit of medical services,
t = own-time input per unit of medical services,
q = money price per unit of X,
s = own-time input per unit of X,
w = earnings per hour,
Y = total (full) income,
y = nonearned income, and
T = total amount of time available for market and own production of goods and services.

---

[4] Similar models can be found in Becker (1965), Grossman (1972*a*), and Holtman (1972).

A more complicated specification of the model was not used because this simpler formulation yields most of the same predictions and because the data do not permit estimation of the unique implications of the richer specification.[5]

Effects of a Change in Price

It can be shown that the assumptions sufficient to make money function as a price in determining the demand for medical services are also sufficient to make time function as a price, producing negative own-time–price elasticities of demand and positive cross-time–price elasticities. One of the chief interests in this study is the relative importance of money prices and time prices in determining the demand for medical services. If we let Π equal the total price per unit of medical services (i.e., $\Pi = p + wt$), then the elasticity of demand for medical services with respect to money price is

$$\eta_{mp} = \frac{p}{\Pi}\eta_{m\Pi}, \tag{2a}$$

and the elasticity with respect to time (which equals the elasticity with respect to wt) is[6]

$$\eta_{mt} = \frac{wt}{\Pi}\eta_{m\Pi}. \tag{2b}$$

Comparing these two elasticities yields the second prediction from the formal model, namely, that $\eta_{mt} \gtreqless \eta_{mp}$ as $wt \gtreqless p$. Clearly, as p goes to zero and wt does not, the time-price elasticity will exceed the money-price elasticity. In other words, as the out-of-pocket payment for a unit of medical services falls, because of either increasing insurance coverage or the availability of subsidized care, demand becomes relatively more sensitive to changes in time prices. Further, this implies that the demand for free medical services should be more sensitive to changes in time prices than demand for nonfree services, because time is a greater proportion of the total price at free than at nonfree providers.

---

[5] Grossman (1972*a*) allows the amount of medical services, m, to influence the total amount of time, T, available for production. Phelps (1973) makes the selection of insurance parameters, and therefore p, endogenous to the system. Some researchers (notably Grossman [1972*a*, 1972*b*] and his followers) have taken the Lancaster (1966) formulation of letting the argument "health" enter the utility function and then deriving a demand for medical services. The present data do not allow us to estimate the relation transforming medical services into health.

[6] These elasticities are only approximate in the long run if insurance premiums are adjusted to reflect the changes in utilization.

Effects of a Change in Income

Exogenous changes in income can arise either from a change in earnings per hour or from a change in nonearned income. The two effects are not, in general, equal. The assumptions that are sufficient to make money function as a price are also sufficient to mean that an increase in nonearned income will produce an increase in the demand for medical services.

The effects of a change in the wage rate cannot be determined a priori because of offsetting influences. An increase in earnings per hour produces an income effect, which acts to increase demand. It also raises the opportunity cost of time, which reduces demand for time-intensive activities. The net effect on the demand for medical services depends on the time intensity of the price of medical services relative to the time intensity of the price of all other goods and services. It can easily be established (see Acton 1973*a* or 1973*b*) that the substitution effect of a change in the wage rate on the demand for medical services is positive if and only if

$$\frac{ws}{(q + ws)} > \frac{wt}{(p + wt)}, \tag{3}$$

that is, if the time price is a larger proportion of the total price for the composite good, X, than it is for medical services, m. The substitution effect is necessarily negative for free sources of medical care since the condition in equation (3) will not be met as long as there is a nonzero monetary price for X. Of course, the net effect of a change in wages may still be to increase the demand for medical services if the income effect exceeds the substitution effect. Intuitively, however, the effect of a wage change on the demand for medical services is more like a price effect for free sources of care (and therefore more likely to be negative) and more like an income effect for nonfree sources of care (and therefore more likely to be positive).

Predictions from Other Formal Models

As noted above, the model developed here is deliberately simplified because it is adequate to produce testable hypotheses for the variables of primary interest. There are some additional hypotheses regarding the effects of education and age from the Grossman (1972*a*) investment model that can be tested with these data. Grossman argues that medical services are combined with other inputs the individual supplies to produce health and that it is health that enters the utility function. Now, if education raises health productivity (e.g., more highly educated people are more skillful in combining inputs to produce health) and if the price elasticity of health is less than one, then education will have a negative effect on the demand for medical care.[7] The second implication of Grossman's work involves the effect of age on consumption of health services. If the price elasticity of demand for health is less than one, then the demand for medical services will be positively correlated with the depreciation rate on health. In general, empirical evidence suggests that the depreciation rate increases over the life cycle, causing a positive effect of age on consumption of medical care.

III. The Empirical Base

The data used for this study come from a 1965 survey of users of the outpatient departments of New York City municipal hospitals. Respondents were selected from a random sample of persons at the clinic; hence, the probability of being interviewed is proportional to the frequency of use of the clinics. The Appendix defines and presents the mean values of the variables used in this analysis. In addition, the means are reported when observations are weighted by the inverse of the number of clinic visits. These weighted means indicate the mean characteristics of people who ever use the clinics, rather than the mean characteristics of the patient loads at the clinics.

The respondents were questioned about their previous year's medical use and a number of sociodemographic characteristics from the previous year. The interviews at each of the OPDs were conducted in four waves spread throughout the year to cancel seasonality in usage.[8] There are advantages and disadvantages to using survey data. One of the advantages is that disaggregated data provide a more precise description of individual behavior because individual rather than aggregate values can be used for explanatory variables. This overcomes the bias away from zero that is frequently encountered in using average values in aggregate data.[9] Other advantages include the much larger sample size typically available in surveys so that variance of coefficients can be reduced. The chief disadvantage of survey data in general is that it relies on recall by the individual. This frequently leads to an underreporting of some variables, particularly medical utilization and income.[10]

---

[7] In the consumption model, given a "neutral" effect of education on all household activities, the elasticity with respect to wealth must also be less than one to produce a negative relationship between medical care and education (Grossman 1972*b*, pp. 36–37).

[8] Details of the study and selected analysis (chiefly analysis of variance) can be found in Lerner and Kirchner (1967) and Lerner, Kirchner, and Dieckmann (1967, 1968).

[9] See Newhouse and Phelps (1974) for an elaboration of this point.

[10] Recall for ambulatory care seems to be extremely accurate for physician use in the last 2 weeks and for hospital use in the last year, according to Regina Loewenstein (personal communication, 1971). Underreporting will generally bias the coefficients of explanatory variables toward zero. The estimated elasticities, however, need not be

Some responses to the surveys were checked independently, increasing the validity of the data, while others were coded poorly for present analysis. The number of outpatient department visits (OPD) and hospitalizations (HOSP) in the preceding year were verified with providers. The distance (DIST) to the hospital of interview was calculated from the person's home address. The number of private physician visits was coded in intervals and had to be calculated. This calculation and all important variables are discussed in detail in Acton (1973*b*).

IV. Estimation Results

The demand for health care by type of provider is estimated from a simultaneous-equation system using two-stage least squares. Four structural equations are specified and 28 exogenous variables are used for estimation. All the equations are overidentified by several variables.[11] After a brief overview of the model, this section describes first the structural equations and then the reduced-form equations.

Overview of the Model

The four structural equations describe the volume of ambulatory and inpatient services demanded and the distance traveled to the free ambulatory care.[12] The last year's volume of OPD visits and the number of

---

biased. If all people recall k proportion of their utilization in the previous year, then the estimated elasticity of demand for care with respect to, say, price is

$$\eta_{mp} = \frac{\partial km}{\partial p} \cdot \frac{p}{km} = \frac{\partial m}{\partial p} \cdot \frac{p}{m},$$

which is the same as the elasticity that would have been estimated with full recall. By the same argument, the cross-price elasticities and the elasticity of substitution of one type of provider for another should be unbiased. There is further potential bias in this particular study because of the fee structure at the municipal hospitals. In order to receive free care, the individual was supposed to be unable to pay for private care (although I am told that it was well known at the time that anyone who asserted he was entitled to free care would receive it without challenge). This institution, however, might have caused some persons to underreport income or utilization of private facilities. There is no evidence that such an error was introduced and the interview opened with a statement of confidentiality, but in the absense of verified income data the possibility exists.

[11] Since a number of the variables are dummy variables for health status or mode of transportation, they really contain less information than their number indicates. Therefore, in checking the number of excluded exogenous variables for purpose of identification, I counted a set of mutually dependent dummy variables as one variable.

[12] Some of the "exogenous" variables in this cross-sectional survey are really endogenous in a larger system that includes life-cycle behavior (such as labor-force participation and family size) and a broader set of economic decisions. Since we are specifying only a subset of these relations, it is possible that some of these exogenous variables are really proxies for common underlying theoretical variables. I tried to limit the possibility of such undesirable interference by specifying such variables as family health status only in the equations where their principal effects could be anticipated.

visits to a private physician (PRIVE) are the two alternative sources of ambulatory care. They are technical substitutes for one another. The volume of private-physician visits and the number of hospitalizations are entered as right-hand endogenous variables in order to measure substitution and complementarity effects among different types of providers. We would prefer to enter the money price of private-physician visits and the cost of a hospitalization and examine the cross-price effects, but this information is not available in the survey. Since substitution or complementarity is important both theoretically and empirically in medical care, we wish to measure the effects as best possible and limit the bias in the remaining coefficients. One consequence of entering quantities on the right-hand side is to make price elasticities different from the usual elasticities, as discussed next.

The distance (DIST) to the outpatient department where the interview was conducted is specified as an endogenous variable in this system. It is the chief measure of the price of an OPD visit and functions as a cross-price term for PRIVE and HOSP.[13] Distance actually measures several things. It includes the physical distance one has to travel and the money and time costs of travel, and it is associated with higher informational costs. The informational costs represent the fact that patients generally have less difficulty in finding out about the quality and suitability of a close-by clinic (for instance, by asking neighbors or by having experienced the care themselves) than in finding out about a distant clinic. The money costs and informational costs will tend to be positively correlated with the distance traveled. However, those who previously lived near one clinic and now live farther away (perhaps nearer another) have a lower information cost with the former, more distant, clinic. Therefore, the coefficient on HABIT should be positive in the structural equation for distance, but it will capture only part of the differences in information costs in this and other equations. DIST is endogenous in this system because the distance a person travels to the OPD is influenced by a number of the same underlying variables influencing demand for medical care, including the opportunity cost of time, the health status of the user, and his sociodemographic characteristics. Consequently, the only way to see the total effect of these variables is through the reduced-form coefficients. The result of this specification is that the elasticities with respect to DIST are not identical with the usual price elasticity of demand because they show the effect of distance holding constant the level of all other things, including other forms of medical care. With three or more goods, it is not possible to sign bias without stronger assumptions about the utility function.[14]

---

[13] *An imprecise measure of waiting time is included. If the person says he waits longer* at an OPD than at a private physician's office WAIT take the value one.

[14] With only two goods, the own-cross-price elasticities estimated holding constant the quantity of the other good are lower bounds (in absolute value) for the normal

The important exogenous economic variables in this system are earned and nonearned income (GRWAGE and NWAGE), assets (LIQUID and EQUITY), and health insurance (INSF). The insurance variable is very imprecise and indicates only that at least one member of the family had some coverage (this was before Medicare and Medicaid). Additional important variables in this system include the mode of transportation used for the named sources of care (the suffix letter *P* indicates mode of private care), health status, and a number of sociodemographic variables.[15]

The Structural Equations

The specifications of the structural equations and their estimated coefficients are given in table 1.[16] In the absence of an explicit utility function, there is no unique set of structural equations. A linear specification is employed. The principal value of the structural equations is to show the effects of endogenous variables on one another. The net effect of exogenous variables is clear only from the reduced-form coefficients in table 2.

The OPD Equation

The structural equation for OPD visits (1) includes the other three endogenous variables. Private-physician visits are technical substitutes for OPD visits, producing an elasticity of −.11. Hospitalizations appear to be significant technical complements to OPD care with an elasticity of .47. This complementarity is reasonable, first, because inpatient care frequently requires ambulatory follow-up. Second, hospitalizations

---

elasticities. With three or more goods, we need stronger assumptions (such as, hospitalizations are complementary to one form of ambulatory care and substitutes to the other form) because of offsetting effects to determine the sign of the bias unambiguously.

[15] It can be argued that the method of transportation is an endogenous variable just as is the distance traveled to OPD care. For both conceptual and practical reasons, it is considered exogenous in this model. First, it can be viewed as exogenous if before the period of observation the person has already made a decision about the methods of transportation he will use for various types of trips. We can argue that he is unlikely to alter this choice significantly during the course of the year. A more practical reason is that making the method of transportation endogenous means that the normal assumptions about the distribution of the error term would not be satisfied, and either a Tobit or probit model would be more appropriate. Estimating a simultaneous Tobit system with 12 or so endogenous variables is probably unwarranted with this data base.

[16] The *t*-values reported are the asymptotically normal values. By multiplying these by the *t*-adjustment factor, the Dhrymes (1969) *t*-statistics for 2SLS result.

TABLE 1

COEFFICIENTS OF STRUCTURAL EQUATIONS FROM 2SLS ESTIMATION*

VARIABLES	OPD (EQ. [1])			PRIVE (EQ. [2])			HOSP (EQ. [3])			DIST (EQ. [4])		
	Coef.	t	η	Coef.	t	η	Coef.	t	η	Coef.	t	η
Endogenous:												
OPD	...	...	...	−0.052	3.45	−.27	0.006	1.58	.14	0.021	0.36	.08
PRIVE	−0.577	3.20	−.11	...	...	...	0.023	2.80	.10	...	...	...
HOSP	11.144	5.22	.47	1.312	2.69	.29	...	...	...	...	...	...
DIST	−0.552	2.51	−.14	0.050	0.99	.07	0.029	2.40	.18	...	...	...
Exogenous:												
GRWAGE × 10^{-4}	1.010	1.02	.04	0.452	2.02	.09	−0.098	1.80	−.09	0.117	0.76	.017
NWAGE × 10^{-3}	0.329	1.47	.04	0.171	3.40	.10	−0.035	2.94	−.10	0.052	1.47	.02
LIQUID × 10^{-4}	−0.161	0.14	−.00	0.419	1.65	.01	0.037	0.58	.00	...	...	...
EQUITY × 10^{-3}	−7.047	0.52	−.02	0.047	1.56	.01	0.014	1.91	.01	...	...	...
INSF	−0.347	0.75	−.02	0.285	2.76	.07	0.054	2.17	.06	...	...	...
WAIT	0.401	0.78	.01	1.174	12.16	.23	...	...	...	−0.049	0.61	−.01
FAMH	0.972	2.30	.04	...	...	...	...	...	...	...	...	...
EX	−1.726	1.90	−.02	−1.406	7.14	−.09	−0.101	2.08	−.03	0.263	1.45	.01
GOOD	−1.083	1.65	−.07	−1.152	8.23	−.36	−0.075	2.15	−.11	0.285	2.29	.07

FAIR	−0.788	1.30	−.03	−0.880	6.63	−.18	−0.015	0.44	−.01	0.014	0.14	.00
AGE × 10^{-1}	0.811	7.59	.38	0.046	1.74	.11	−0.019	3.05	−.21	0.025	0.70	.04
BLACK	−2.312	4.51	−.12	−0.382	3.19	−.10	0.055	1.87	.07	−0.240	2.20	−.05
PR	−1.643	2.95	−.07	−0.446	3.51	−.10	0.016	0.51	.02	−0.261	2.53	−.05
EDUC	−0.131	2.77	−.11	0.007	0.64	.03	0.005	1.80	.09	...	...	...
MALE	−1.863	4.35	−.09	−0.340	3.47	−.09	0.094	4.26	.11	0.158	2.24	.03
WELF	−0.827	1.50	−.03	−0.411	3.36	−.07	0.064	2.14	.05	...	...	...
WORK	...	...	...	0.251	1.99	−.03	−0.059	1.92	−.03	0.095	0.69	.01
NMAIN	−2.372	2.48	−.01	...	...	...	...	...	...	0.505	3.10	.01
HABIT	0.137	4.72	.09	...	...	...	...	...	...	−0.004	0.54	−.01
BUS1	...	...	...	...	...	...	...	...	...	1.067	12.35	.23
BUS2	...	...	...	...	...	...	...	...	...	2.444	25.01	.28
CAR	...	...	...	...	...	...	...	...	...	2.524	14.72	.06
TAXI	...	...	...	...	...	...	...	...	...	1.393	10.85	.06
BUS1P	...	...	...	2.986	17.49	.13	...	...	...	...	...	...
BUS2P	...	...	...	3.193	13.75	.74	...	...	...	...	...	...
CARP	...	...	...	2.125	7.70	.47	...	...	...	...	...	...
TAXIP	...	...	...	2.958	9.95	.43	...	...	...	...	...	...
CLINS	...	...	...	...	...	...	...	...	...	−0.099	−0.71	−.09
CONSTANT	5.636	4.30	...	1.363	4.47	...	0.240	3.50	...	0.515	1.80	...
t-adjustment factor	0.19	...	...	0.55	...	...	0.347	...	...	0.763	...	...

* Scaling factors adjust only *coefficients*, not elasticities. Elasticities calculated at the means of unweighted data.

TABLE 2

Coefficients of Reduced-Form Equations*

Variable	OPD (Eq. [5]) Coef.	OPD (Eq. [5]) η	PRIVE (Eq. [6]) Coef.	PRIVE (Eq. [6]) η	HOSP (Eq. [7]) Coef.	HOSP (Eq. [7]) η	DIST (Eq. [8]) Coef.	DIST (Eq. [8]) η
GRWAGE × 10^{-4}	−0.245	−.01	0.353	.07	−0.089	−.08	0.111	.02
NWAGE × 10^{-3}	−0.129	−.01	0.139	.08	−0.031	−.09	0.049	.02
LIQUID × 10^{-4}	0.100	.00	0.477	.01	0.048	.00	0.002	.00
EQUITY × 10^{-3}	0.070	.00	0.064	.01	0.016	.01	0.001	.00
INSF	0.150	.01	0.360	.09	0.063	.07	0.003	.00
WAIT	0.025	.00	1.205	.23	0.026	.02	−0.048	−.01
FAMH	1.050	.05	−0.046	−.01	0.006	.01	0.022	.00
EX	−2.600	−.03	−1.450	−.09	−0.143	−.04	0.208	.01
GOOD	−1.709	−.10	−1.190	−.37	−0.106	−.15	0.249	.06
FAIR	−0.713	−.03	−0.895	−.18	−0.040	−.04	−0.0003	−.00
AGE × 10^{-1}	0.631	.30	−0.0003	−.01	−0.014	−.16	0.037	.06
BLACK	−1.661	−.08	−0.269	−.07	0.031	.04	−0.275	−.05
PR	−1.354	−.06	−0.402	−.09	−0.009	−.01	−0.290	−.05
EDUC	−0.089	−.07	0.018	.08	0.005	.09	−0.002	−.01
MALE	−0.854	−.04	−0.171	−.04	0.090	.10	0.140	.03
WELF	−0.0004	−.00	−0.337	−.05	0.057	.04	−0.000	−.00
WORK	−0.805	−.02	0.222	.02	−0.057	−.03	0.078	.01
NMAIN	−2.705	−.02	0.167	.01	0.001	.00	0.449	.01
HABIT	0.149	.10	−0.007	−.02	0.001	.01	−0.0004	−.01
BUS1	−0.299	−.02	0.110	.03	0.031	.04	1.063	.23
BUS2	−0.683	−.02	0.251	.04	0.072	.05	2.429	.28
CAR	−0.706	−.00	0.259	.01	0.074	.01	2.509	.06
TAXI	−0.390	−.00	0.143	.01	0.041	.01	1.385	.06
BUS1P	−1.079	−.01	3.125	.14	0.065	.01	−0.023	−.01
BUS2P	−1.148	−.01	3.342	.08	0.070	.01	−0.024	−.00
CARP	−0.763	−.00	2.224	.05	0.046	.00	−0.016	−.00
TAXIP	−1.063	−.00	3.095	.05	0.064	.00	−0.022	−.00
CLINS	0.027	.01	−0.010	−.01	−0.002	−.01	−0.099	−.08
CONSTANT	8.237	...	1.412	...	0.340	...	0.687	...

* Scaling factors adjust only coefficients, not elasticities. Elasticities calculated at the means of unweighted data.

produce a budget effect that may increase the demand for less expensive forms of ambulatory care. Third, the municipal OPD clinics act as ports of entry to municipal hospitals.

One of the important predictions of the formal model developed above is that distance to the OPD would function as a price in determining demand. Equation (1) supports this prediction, producing an elasticity of $-.14$. It is likely that the elasticity with respect to travel time is greater (in absolute value) than $-.14$ because people traveling farther will tend to take more rapid forms of transportation.[17] The mode of transportation to OPD care has its major influence on the distance traveled and is included in the DIST structural equation.

The PRIVE Equation

The structural equation for private physician care complements the estimates from the OPD equation. Again, PRIVE and OPD are substitutes and PRIVE and HOSP are complements. DIST functions as a cross price to PRIVE, producing an elasticity of about .07. The estimate of .07 is considerably smaller than the cross-time elasticities of about .6 estimated for private care in Acton (1973*a*). Again, it is likely that the elasticity with respect to travel time is greater. An additional downward bias in the coefficient may be caused by the omission of a measure of travel distance from the structural equation for PRIVE. All we have to measure travel distance are the dummy variables for method of transportation. In this structural equation, the partial effects indicate that walkers demand the least care, while taking the bus, subway, or taxi all have about the same effect. The statistically significant coefficients on mode of transportation suggest that distance is playing an important role in demand (this conclusion is supported by the reduced-form coefficients reported in table 2).

The Hospitalization Equation

The structural equations for hospitalization reveal effects for the endogenous variables consistent with the other equations. Clinic and private-physician visits are complements to hospitalization, with elasticities around .14 and .10. The coefficient on distance to an OPD has a positive sign (elasticity = .18)

---

[17] For instance, it may not be worth taking a bus a few blocks, but it is worthwhile when crossing town; similarly, it may be worthwhile to switch to an express train for part of a longer journey by subway and then return to a local. The own-price elasticity of demand for public ambulatory care with respect to self-reported travel time is estimated between $-.6$ and -1.0 in Acton (1973*a*).

The Distance Equation

The only included endogenous variable in explaining distance to the OPD is volume of OPD services, arguing that those who go more frequently will want to reduce the travel price they have to pay. The results of equation (4) show a coefficient not statistically significantly different from zero. The distance equation is specified as a function of a number of the exogenous variables that are common to the medical care equations—including earned and nonearned income, mode of transportation, health status, and sociodemographic characteristics. The most important theoretical prediction for this equation is that those with a higher opportunity cost of time should travel shorter distances for OPD care. The coefficient on GRWAGE in equation (4) does not support this prediction, although its *t*-ratio is only 0.76. The general effects of the opportunity cost of time follow the predictions as indicated by the coefficient on GRWAGE in the reduced-form equations for medical care.

The Reduced-Form Equations

The chief value of the reduced-form equations is to indicate the net influence of the exogenous variables on the dependent variables. Consequently, the effects of a few key variables will be examined across equations, and the estimated coefficients will be compared with the predictions of Section II.

The Role of Travel Distance and the Value of Time

There were several predictions about the role of time on the demand for various types of care. The first prediction, that time (and travel distance) would function as a price, was supported by the structural equations. They reveal a negative (own-price) coefficient on the variable DIST in the OPD equation and a positive (cross-price) coefficient in the PRIVE and HOSP equations. Second, we expect those with a higher opportunity cost of time to demand less time-intensive care. Both working people and people with a higher opportunity cost of time demand less time-intensive OPD and hospital care and more private-physician care. A further hypothesis, that the higher wage rate would shorten travel distance, was not supported by equation (4).

Income and Assets

Unless service from a provider is an inferior good, the nonearned income elasticity should be positive. Equations (5), (6), and (7) show a positive elasticity of demand for private-physician services and a negative elasticity for OPD and hospital care, suggesting that private outpatient care is a normal good and municipal health care is an inferior good (in an economic sense). The effects of earned income depend on the time intensity of medical goods relative to other goods and services. The model predicted that earned income would function more like a price effect in the demand for free care and more like an income effect in the demand for nonfree care. The reduced-form coefficients show a positive elasticity of demand for private outpatient care with respect to GRWAGE and a negative elasticity of demand for public care. This supports the popular impression that OPD and hospital care are time intensive relative to private physicians' care and relative to all other goods and services.

Education and Age

Two other predictions from the formal models involve the effects of education and age. Grossman suggested that, if those with more education were more efficient producers of health, education would have a negative coefficient (as long as the price elasticity of demand for health is less than one). Equations (5) and (6) show a negative effect of education on OPD visits and a positive effect on private-physician visits. However, the decrease in the number of OPD visits is only partly made up by increases in the number of private visits so that the net change in ambulatory visits produced by an increase in education is negative. This finding is compatable with the predictions of Grossman's (1972*b*) model discussed above. The investment model also predicted a positive correlation of age and the depreciation rate on health. The present data on ambulatory utilization show a negative (but very small) elasticity for PRIVE and a positive net effect of age for OPD care. The OPD finding supports the hypothesis that depreciation in the health stock increases with age. The negative coefficient on age in the hospitalization equation may indicate older persons going to nursing homes.

The remainder of the coefficients describing health status, insurance, and sociodemographic characteristics of the population generally conform to the expectations based on other studies. As found by most researchers, those in poorer health demand more inpatient and outpatient care. They also travel shorter distances to receive their OPD care. The effect of at least one family member's having fair or poor health (FAMH) reinforces the health status effects. In general, the positive correlation of health status in the family and the budget effect of a very sick member should either increase utilization at each source of care or cause a shift to public care; FAMH is positive in the OPD equation (5). Although it is an imprecise variable, if anyone in the family has health insurance, the person will seek more medical services, both public and private. Blacks and Puerto Ricans generally receive less care of all types than do their white counterparts, and they travel shorter distances to receive it. Finally, men seek less ambulatory care and more inpatient care than women. This finding provides further support for the suggestion of Acton (1973*a*) that males may let their health deteriorate further than females do before seeking medical care, so that when they go they require more intensive care.

V. Conclusion

This study supports the prediction that travel time (as measured by distance) functions as a price in determining the demand for medical services when free care is available. This survey of users of the municipal OPDs indicates negative own-price elasticities with respect to travel distance at free providers and positive cross-price elasticities for nonfree providers of care. Further, the estimated distance elasticity approaches or equals the money-price elasticities that have been estimated by a number of researchers.[18] The predicted negative effect of earned income on distance was not found, but persons with higher earned income are more likely to use the private sector, which is relatively less time intensive, than the public sector.

There are a number of policy implications of this study. Two of the most important involve (1) the redistribution in services that will be caused by a change in money and time prices and (2) the possibility of using income subsidies rather than direct provision of goods to meet public objectives. As money prices out of pocket are reduced, because of either the continued spread of private insurance or the enactment of a federal health insurance scheme, demand will become more responsive to time prices. In turn, this will permit persons with a lower opportunity cost of time to bid services away from those with a higher opportunity cost of time because they will face a price that is relatively less costly. This conclusion holds even if there is not differential coverage by income class and even if there is no supply side response increasing the time prices. It is likely that a shift in demand will be accompanied by an increase in the time needed to receive a unit of medical services.[19] This will further increase the relative shift in favor of those with a lower opportunity cost of time. The significant effect of travel distance on the demand for medical services suggests a policy instrument for delivering more services to target groups. By moving clients closer—either by improving transportation, locating clinics closer, or by establishing satellite clinics around central facilities—the consumption of selected populations can be increased.

---

[18] Feldstein (1971), Davis and Russell (1972), and Rosett and Huang (1973) have all reported money-price elasticities in the range −.5 to −1.0, although there is reason to believe these estimates are biased upward (see Newhouse and Phelps 1974). Other more conservative measures of the money-price elasticity (using several data sources) place it around −0.15 or less (see Phelps and Newhouse 1974).

[19] This supply response is likely for a number of reasons. First, it may be optimal from the point of view of the provider to have a queue to even out the variation in demand that he experiences without having to invest in significant excess capacity. A shift in demand will generally cause the optimal queue length to change (for instance, the opportunity cost of an idle moment of the supplier's facility is higher). Second, the suppliers may not be profit maximizers, so that they do not respond to a shift in demand by charging

A second important policy implication is for alternative means of meeting the objective of increasing medical services consumed by target populations. In one form or another most proposals reduce to a subsidized provision of services, whether through social insurance schemes such as Medicaid or various federal health insurance proposals, or through direct provision of care as in neighborhood health centers or the requirement that Hill-Burton hospitals provide charity care. But as Davis (1972) has correctly noted, there is seldom a consideration of the extent to which changing the income distribution will alleviate the desire to subsidize the medical purchase. Even in the administration's proposals for income maintenance (FAP) and subsidized medical care for the poor (FHIP) there was little discussion of the degree to which one can be substituted for the other. The proposed comprehensive health insurance plan (CHIP) now before Congress provides an opportunity to illustrate the tradeoffs.

Income subsidy will not, in general, meet the objective of risk spreading for medical expenses (unless it induces a significant demand for health insurance), but it will raise the average demand for medical services in the subsidized population. The equations reported in tables 1 and 2 put us in a position to address this question of substituting income maintenance for subsidized medical care to achieve a given increase in medical services consumption. Since income maintenance is a nonearned source of income, the elasticity of demand for medical care with respect to changes in nonearned income is used. A hypothetical example, not based exactly on FAP and CHIP provisions, will serve to illustrate. The estimation results in table 2 indicate that a $1,100 increase in nonearned income for a family with a current nonearned income of about $400 and earned income of about $2,900 (in 1965) will produce an increase of about 11 percent in the demand for private practitioners' care per person. If the money-price elasticity of demand for ambulatory medical services is around −.15 over the range under consideration,[20] and the out-of-pocket

---

the highest possible monetary prices but instead allow time prices to increase. In particular, physicians may be income satisfiers rather than maximizers. See Newhouse (1970), Frech and Ginsburg (1972), and Newhouse and Sloan (1972) for a discussion of physician pricing behavior. Third, there may be a conscious attempt to redistribute services by discriminating in favor of those with a lower opportunity cost of time. See Nichols, Smolensky, and Tideman (1971) for a discussion of the first and third points.

[20] The actual money-price elasticity may be even lower than this. See Phelps and Newhouse (1974) for a discussion of the price elasticities in several published reports.

expenditure is reduced from 25 percent of money price to 15 percent (the upper limit on CHIP's coinsurance rate and the rate for an insuree with income of $3,000), then the demand for private care will increase by 7.5 percent. Clearly, one means of increasing private medical consumption by the poor is income supplementation, and the magnitude of the change may be comparable over the range of subsidy and income guarantee under consideration.

Appendix

Definition of Variables Used and Their Mean Values[21]

AGE = Age in years. Means = 35.6, 31.1.
BLACK = Dummy variable equaling one if Negro or indeterminable, or other than Puerto Rican, Mexican-American, American Indian, or other white. Means = 0.38, 0.41.
BUS1 = Dummy variable equaling one if patient's usual means of transportation to the clinic requires one bus or train. Means = 0.42, 0.42.
BUS2 = Dummy variable equaling one if patient's usual means of transportation to the clinic requires two or more buses or trains. Means = 0.23, 0.21.
BUS1P = Dummy variable equaling one if patient's usual means of transportation to private doctor requires one bus or train. Means = 0.07, 0.08.
BUS2P = Same as above except two or more buses or trains. Means = 0.03, 0.03.
CAR = Dummy variable equaling one if patient's usual means of transportation to the clinic was by car driven either by individual or a friend. Means = 0.05, 0.06.
CARP = Same as above except to private doctor. Means = 0.03, 0.03.
CLINS = Total number of clinics used last year. Means = 1.71, 1.29.
DIST = Distance in miles to the hospital of interview. Means = 2.14, 2.13.
EDUC = Highest grade completed, in years. Means = 6.25, 6.23.
EQUITY = Equity in home. Means = $207, $198.
ERNRS = Number of earners in the family. Means = 0.83, 0.91.
EX = Dummy variable equaling one if health status of patient is excellent. Means = 0.095, 0.13.
FAIR = Dummy variable equaling one if health status of patient is fair. Means = 0.30, 0.27.
FAMH = Dummy variable equaling one if all family members reported health as good. Means = 0.34, 0.33.
GOOD = Dummy variable equaling one if health status of patient is good. Means = 0.46, 0.48.
GRWAGE = Gross annual earnings from all wage earners in the family. Means = $2,929, $3,215.
HABIT = Number of years patient has been coming to current clinic. Means = 5.03, 4.10.

---

[21] The first mean value is for the unweighted data and the second is for the data weighted by 1/OPD to adjust for the probability of being sampled.

HOSP = Number of hospitalizations last year. Means = 0.32, 0.30.
HSSIZE = Number of persons in individual's household. Means = 3.67, 3.92.
INSF = Dummy variable equaling one if any family member has medical insurance. Means = 0.35, 0.36.
LIQUID = Liquid assets. Means = $177, $207.
LNGWT = Dummy variable equaling one if patient had to wait a long time before being taken care of at the hospital where he was interviewed. Means = 0.59, —.
MALE = One if male, zero if female. Means = 0.38, 0.39.
NMAIN = Dummy variable equaling one if main source of medical care is other than the same clinic as at time of interview. Means = 0.047, 0.064.
NWAGE = Nonearned family income in last year. Means = $878, $802.
OPD = Number of visits in last year to a physician in a clinic. Means = 7.65, 2.97.
PR = One if Puerto Rican; zero otherwise. Means = 0.34, 0.35.
PRIVE = Number of visits in last year to a physician in his private office. Means = 1.46, 1.63.
TAXI = Dummy variable equaling one if patient's usual means of transportation to the clinic was by taxi. Means = 0.091, 0.095.
TAXIP = Same as above except to his private doctor. Means = 0.021, 0.020.
WAIT = Dummy variable equaling one if patient had to wait longer in the hospital where he was interviewed than in private doctor's office. Means = 0.28, —.
WELF = Dummy variable equaling one if individual had some type of welfare assistance. Means = 0.24, 0.22.
WORK = One if indivudal worked either full or part time. Means = 0.16, 0.20.

References

Acton, Jan Paul. "Demand for Health Care among the Urban Poor with Special Emphasis on the Role of Time." Memorandum R-1151-OEO/NYC, RAND Corp., April 1973.(*a*)

———. "Demand for Health Care When Time Prices Vary More than Money Prices." Memorandum R-1189-OEO/NYC, RAND Corp., May 1973.(*b*)

Becker, Gary. "A Theory of the Allocation of Time." *Econ. J.* 75 (September 1965): 493–517.

Davis, Karen. "Health Insurance." In *Setting National Priorities: The 1973 Budget*, edited by Charles Schultze et al. Washington, D.C.: Brookings Institution, 1972.

Davis, Karen, and Russell, Louise. "The Substitution of Hospital Care for Inpatient Care." *Rev. Econ. Statis.* 54 (May 1972): 109–20.

Dhrymes, Phoebus. "Alternative Asymptotic Tests of Significance and Related Aspects of 2SLS and 3SLS Estimated Parameters." *Rev. Econ. Studies* 36 (April 1969): 213–26.

Feldstein, Martin S. "An Economic Model of the Medicare System." *Q.J.E.* 85 (February 1971): 1–20.

Frech, H. E., and Ginsburg, Paul B. "Physician Pricing: Monopolistic or Competitive: Comment." *Southern Econ. J.* 38 (April 1972): 573–77.

Grossman, Michael. "On the Concept of Health Capital and the Demand for Health." *J.P.E.* 80, no. 2 (March/April 1972): 223–55.(*a*)

———. *The Demand for Health: A Theoretical and Empirical Investigation.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1972.(*b*)

Harris, Seymour E. "The British Health Experiment: The First Two Years of the National Health Service." *A.E.R.* 41 (May 1951): 652–66.

Holtman, A. G. "Price, Time, and Technology in the Medical Care Market." *J. Human Resources* 7 (Spring 1972): 179–90.

Lancaster, Kevin. "A New Approach to Consumer Theory." *J.P.E.* 74, no. 2 (April 1966): 132–57.

Lerner, Raymond C., and Kirchner, Corinne. *Municipal General Hospital Outpatient Population Study: Social and Economic Characteristics of Patients in New York City Outpatient Departments, 1965: Financial Eligibility under Medicaid and Potential Reimbursement to the City.* Report no. 1. New York: School Public Health and Admin. Medicine, Columbia Univ., 1967.

Lerner, Raymond C.; Kirchner, Corinne; and Dieckmann, Emil. *Municipal General Hospital Outpatient Population Study: Social and Economic Characteristics of Patients in New York City Outpatient Departments, 1965: Methodology.* Report no. 2. New York: School Public Health and Admin. Medicine, Columbia Univ., 1967.

———. *New York Municipal General Hospital Outpatient Population Study, 1965: Data on Background, Medical Care Utilization and Attitudes of Outpatients, by Hospital.* Report no. 3. New York: School Public Health and Admin. Medicine, Columbia Univ., 1968.

Leveson, Irving. "Demand for Neighborhood Medical Care." *Inquiry* 7 (December 1970): 17–24.

Newhouse, Joseph P. "A Model of Physician Pricing." *Southern Econ. J.* 37 (October 1970): 174–83.

Newhouse, Joseph P., and Phelps, Charles E. "On Having Your Cake and Eating It Too: A Review of Estimated Effects of Insurance on Demand for Health Services." Memorandum R-1149-NC, RAND Corp., April 1974.

Newhouse, Joseph P., and Sloan, Frank A. "Physican Pricing, Monopolistic or Competitive: Reply." *Southern Econ. J.* 38 (April 1972): 577–80.

Nichols, D.; Smolensky, E.; and Tideman, T. N. "Discrimination by Waiting Time in Merit Goods." *A.E.R.* 61, no. 3, pt. 1 (June 1971): 312–23.

Phelps, Charles E. "Demand for Health Insurance: A Theoretical and Empirical Investigation." Memorandum R-1054-OEO, RAND Corp., July 1973.

Phelps, Charles E., and Newhouse, Joseph. "The Effects of Coinsurance on Demand for Physician Services." Memorandum R-976-OEO, RAND Corp., June 1972.

———. "Coinsurance, the Price of Time, and the Demand for Medical Services." *Rev. Econ. and Statis.* 56, no. 3 (August 1974): 334–42.

Rosett, Richard N., and Huang, Lien-fu. "The Effects of Health Insurance on the Demand for Medical Care." *J.P.E.* 81, no. 2 (March/April 1973): 281–305.

Wealth Effect on the Demand for Money

Vikas Chitre

University of Jodhpur, India

The wealth effect on the individual's money demand is examined within the inventory theoretical framework involving stochastic cash needs. A positive wealth effect arises if an increase in wealth increases the individual's planned expenditure beyond his expected receipts, or if the marginal yield on earning assets is diminishing or marginal penalty cost is increasing. The *direct* wealth effect due to decreasing absolute risk aversion is negative rather than ambiguous as obtained by Arrow (1971). Finally, interpreting penalty cost as the utility lost due to inadequate money holding could give rise to a positive wealth effect, but the presence of near-money assets makes it negative.

Introduction

It has long been recognized that the transactions and precautionary demand for money can be analyzed within the framework of general inventory problems (Baumol 1952; Tobin 1956; Arrow, Karlin, and Scarf 1958; Miller and Orr 1966; Tsiang 1969). The purpose of this paper is to examine, within the inventory theoretic framework, the role played by the individual's wealth in determining his demand for money and to analyze the effect of changes in his wealth on his optimum cash balance.

Why should a measure of wealth enter the individual's demand for money? One explanation could be that wealth determines the rate of current expenditures of the individual, as in Ando and Modigliani (1963), and the rate of current expenditures in turn determines the cash balances

---

I am grateful to Geoff Harcourt of the University of Adelaide, Australia, and two anonymous referees of this *Journal* for helpful comments.

[*Journal of Political Economy*, 1975, vol. 83, no. 3]

© 1975 by The University of Chicago. All rights reserved.

required for transactions and precautionary motives.[1] However, this is hardly the argument made by the proponents of the so-called wealth approach to the demand for money. Their argument, briefly spelled out in Friedman (1956) and Turvey (1960), runs as follows. Ultimate wealth-owning units are assumed to have a scale of preferences among different combinations of assets depending upon the flow of services yielded by them. Each wealth-owning unit is then viewed as choosing, subject to the wealth constraint, the asset composition which maximizes the "utility" of the stream of services flowing from different forms of holding wealth. Money being one of the assets, the optimal money holdings are also subject to the wealth constraint. Thus some measure of wealth would enter as an argument of the individual's demand function for money. Since money is "interpreted as claims or commodity units that are generally accepted in payments of debts at a fixed nominal value" (Friedman 1956, p. 5), the services of money must be considered in terms of the extent to which its ready availability enables the individual to make prompt payments and to avoid embarrassment or losses associated with numerous, inopportune, or hasty transfers of other assets into cash or with expensive borrowing arrangements. Our approach, using ideas from the inventory models, enables us to characterize more precisely the factors on which the services of money holdings depend. By explicitly introducing the constraint of wealth on the individual's money holdings in our model, therefore, we shall be able to examine more clearly the theoretical argument of the wealth approach to the demand for money.

The effect of wealth on the money demand at the micro level has been a subject of some theoretical discussion in recent years. Arrow (1971, chap. 3) has argued, for example, that if the individual's relative risk aversion increases with his wealth, the wealth elasticity of money demand must be at least unity. If the relative risk aversion of the individual increases with wealth, this implies that with an increase in wealth, the individual is increasingly more unwilling to take a bet equal to a given proportion of his wealth or, in other words, to hold a given proportion of his wealth in risky assets. Although one can understand how this argument might explain a wealth elasticity of larger than unity for secure assets or for liquid assets, it is not clear how this argument can be used to justify the observed wealth elasticity of money according to the narrow definition of money as currency plus demand deposits.

---

[1] We may exclude the speculative motive on the argument that in the present-day institutional setup, money as a store of value is likely to be dominated by risk-free near-money assets with a positive yield (see, e.g., Teigen 1964, pp. 482–83; Tsiang 1969, p. 100). It is of course possible that the interest on near-money assets might not be adequate to cover the transactions costs on them for small quantities of speculative balances or speculative balances held for a short period of time. In this case, these speculative balances will be held in the form of narrow money until they grow sufficiently to make their investment in near-money assets profitable. But if we adopt the distinction made by Feige (1967) between the equilibrium money balances and the disequilibrium money balances, the latter being due to partial adjustment on account of the presence of transactions costs, then we can argue that in the *equilibrium* portfolios of individuals, there should be no money balances held for the speculative purpose.

In terms of the narrow definition of money and in terms of an inventory theoretic analysis of transactions and precautionary demand for money, the question of the wealth effect in the money demand function was also briefly reopened by Tsiang (1969, p. 115), and it deserves closer attention. Our paper treats this question more fully and presents several aspects of the problem not examined in Tsiang's paper.

In this paper, we shall use the inventory theoretic framework (see, e.g., Arrow et al. 1958; Tsiang 1969) to examine the role of wealth in influencing an individual's optimum cash balance. The view taken here is that an individual economic agent with a given total (human and nonhuman) net wealth determines his optimal cash balance so as to minimize the expected cost of meeting imperfectly foreseen net financial claims on it, or more generally, so as to maximize the expected utility of his final (or end of period) net wealth when current receipts and payments are imperfectly foreseen and when any consequent inadequacy of cash balance has to be met by incurring a shortage or penalty cost. The gross penalty cost consists of the higher cost of making quick and inopportune changes in the portfolio of earning assets or that of making more expensive borrowing arrangements, etc. This paper thus focuses attention on the money, near-money margin of the individual. The total combined holdings of money and near-money assets of negligible risk of capital loss may be assumed to be determined together *vis à vis* the risky assets by consideration of the individual's behavior toward the risk of future changes in the level of interest rates.

The rest of this paper is divided into five sections. In Section I we show the wealth effect on money demand as arising through its effect on the expected receipts and on the expected payments of the individual. In Section II, we introduce a constraint on the individual's capacity to borrow, termed his line of credit, and examine the likely indirect effect of wealth on money demand through its effect on the individual's line of credit. The relevance to the discussion of the wealth effect on money demand of capital market imperfection and increasing marginal penalty cost is brought out. In Sections I, II, and III, we ignore the consideration of the individual's risk aversion. This implies the use of a linear utility function of the end-of-period wealth and reduces the determination of the individual's optimum cash balance to one of minimizing the expected cost of meeting imperfectly foreseen financial claims on him. In Section IV, we introduce more general utility functions and analyze the effect of wealth on an individual's money demand when his absolute risk aversion is assumed to be decreasing with his wealth. As has been emphasized by Arrow (1971), this assumption rather than the one of increasing absolute risk aversion is the more realistic one to make because the assumption of increasing absolute risk aversion has the absurd implication that with an increase in wealth the individual becomes more unwilling to take a fair bet involving a fixed amount. It is shown in Section IV of our paper that in the context of the inventory theoretic framework, the wealth effect on money demand is perverse (i.e., negative) if the more realistic assumption of decreasing absolute risk aversion is made.

If the marginal penalty cost increases with an increase in wealth, this would lead to another source of a wealth effect on an individual's money demand. A natural way in which an increase in wealth could lead to an increase in the marginal penalty cost is the following. Suppose that not having an adequate money balance requires the individual to postpone expenditures in excess of his present money holding; then the marginal penalty cost may be identified with the excess of the marginal utility of present consumption over the (discounted) marginal utility of future consumption. With a diminishing marginal utility of one-period consumption expenditure, the marginal penalty cost in the above sense increases with wealth. This influence of the wealth variable on money demand is examined in Section V. In this context, it is shown that the presence of near-money assets makes the wealth effect on money demand negative, although in the absence of the near-money assets, it could be positive.

The upshot of the discussion is the following. Because of the presence of multicollinearity, a clear-cut empirical isolation of the relative roles played by a measure of the wealth variable and a measure of the current planned expenditure in the money demand function may be difficult. However, our discussion suggests that *in the absence of the indirect effect through planned expenditure*, the role played by wealth in money demand functions is not likely to be significant.

I

Let W = the total initial net wealth of the individual and M = the individual's money balance. We shall assume that the marginal expected yield over the unit period on the earning assets is constant and equal to r, so that $r \cdot (W - M)$ = the expected yield over the unit period from the earning assets; also P = the expected expenditure of the individual over the unit period.

The excess payments due (that is, accumulated excess of payments over receipts) during the unit period will be assumed to be a random variable given by

$$x = P - r(W - M) + v, \qquad (1)$$

with $E(v) = 0$. The probability density functions of x and v will be denoted by $f(x)$ and $g(v)$, respectively, and their cumulative distribution functions by $F(x)$ and $G(v)$, respectively. For convenience, we shall let

$$S = r(W - M) - P. \tag{2}$$

The gross penalty cost, giving the cost of liquidating, or borrowing against, the earning assets will be denoted by $H(\cdot)$, and it will be a function of the amount of cash deficiency. We shall generally assume $H(0) = 0$, $H'(0) > r$, and $H''(\cdot) \geq 0$.

The individual's optimization problem may be considered here to be one of determining the cash balance which maximizes his expected final wealth:

$$E = W + S - \int_{M}^{\infty} H(x - M) f(x)\, dx. \tag{3}$$

Or, making a change of variable from x to $v = x + S$, we get

$$E = W + S - \int_{M+S}^{\infty} H(v - M - S) g(v)\, dv. \tag{4}$$

Differentiating (4) partially with respect to M on the assumption that $H(0) = 0$,[2] and keeping in mind from the definition of S given in (2) that $\partial S/\partial M = -r$, we obtain the expected marginal return from holding money:

$$E_M = -r + (1 - r) \int_{M+S}^{\infty} H'(v - M - S) g(v)\, dv. \tag{5}$$

The expected marginal return from holding money as given by equation (5) is to be interpreted as the saving at the margin in the expected penalty costs over and above the interest foregone at the margin. In equation (5) the saving at the margin in the expected penalty costs is $(1 - r)$ times the expected marginal penalty cost because an increase in the money holding also has the effect of lowering the expected receipts in proportion to the marginal expected yield on the earning assets.

The optimum cash balance of the individual is obtained by setting

---

[2] This means that we are ignoring the fixed component of the penalty cost here. Even when the fixed component of the penalty cost is present, the conclusions of this paper are not affected as long as the expected marginal utility is assumed to diminish with the individual's money holding. If, as his wealth increases, the individual puts a higher evaluation on the time spent in making the required portfolio adjustments or the borrowing arrangements, this will be an additional source of wealth effect on money demand. However, the fixed component of the penalty cost in this sense should more appropriately be assumed to vary with the individual's income than with his wealth.

$E_M = 0$. The expected marginal return from money holding is easily seen to be diminishing with the size of the individual's money balance, as

$$E_{MM} = -(1-r)^2 \int_{M+S}^{\infty} H''(v-M-S)g(v)\,dv - (1-r)^2 H'(0)g(M+S) < 0.^{3} \tag{6}$$

The effect of a change in the individual's (initial) wealth on his money holding can be found by differentiating (5) partially with respect to W to obtain

$$E_{MW} = -(1-r)r \int_{M+S}^{\infty} H''(v-M-S)g(v)\,dv - (1-r)rH'(0)g(M+S) < 0. \tag{7}$$

Equation (7) shows that an increase in the individual's wealth by increasing his expected receipts would actually lower the expected marginal return on holding money and hence an increase in the individual's wealth would *reduce* the individual's demand for money.[4]

Equation (7) ignores the effect of wealth on money demand through its effect on the expected level of expenditure, P. If this indirect effect of wealth on money demand is taken into account, equation (7) will be modified to become

$$E_{MW} = (1-r)[P'(W)-r] \int_{M+S}^{\infty} H''(v-M-S)g(v)\,dv + (1-r)[P'(W)-r]H'(0)g(M+S). \tag{8}$$

---

[3] The reader will notice that if $H'' = 0$, then the expected marginal return from money holding is strictly diminishing and E, the expected final wealth, is a concave function of M only if $g(M+S) > 0$ for all M. If $g(M+S) = 0$ for some $M > 0$, which will be the case, for example, if g (or f) is a discrete distribution, then E is a concave function of M consisting of linear segments.

[4] One possible source of positive wealth effect on the demand for money in the present context would be if the marginal yield on earning assets is taken to be diminishing. If the marginal yield on earning assets is diminishing, a larger wealth will mean a smaller marginal opportunity cost of holding money and hence will tend to increase the individual's optimum money balance. Thus wealth would have a positive effect on money demand for this reason in simple, money real-capital models. However, in a model involving near-money assets, the marginal yield on earning assets is likely to be the market rate of yield on near-money assets and may be taken to be constant for all practical purposes, as the market rates of yield on near-money assets are most unlikely to fall when an individual wealth owner is slightly richer. However, one possible explanation of the observed wealth effect on money demand, so far overlooked in the literature, might be that an increase in the community's *aggregate* wealth increases the community's demand for near-money assets, reducing the yield on money substitutes. This, in turn, increases money (narrowly defined) which individuals wish to hold. Thus there could be an *indirect* effect of *aggregate* wealth on *aggregate* money demand through changes in the interest rates on near-money assets.

Thus if the individual's expected payments increased with wealth more than his expected receipts did, then this would be a source of positive wealth effect on his demand for money. Clearly, if the effect of an increase in wealth on current planned expenditure is ignored, the observed wealth effect on money demand cannot be explained in terms of the model presented in this section.

II

In the previous section we did not impose any restriction on the amount of borrowing that the individual can resort to. The maximum amount of borrowing which the individual can possibly avail himself of should normally depend upon his net wealth. We might then investigate the effect of an increase in wealth on the individual's demand for money when the constraint on his borrowing ability is explicitly introduced.

When the individual's cash holding is not adequate to meet his current excess payments due, we shall assume that he can liquidate, or borrow against his total net wealth. Given the structure of the individual's assets (for example, their division between liquid and nonliquid assets), the maximum adjustment which the individual can effect, *inclusive* of that made from his own cash balance, depends upon his net wealth, and we shall assume it to be given by

$$L = M + \theta[W - M], \tag{9}$$

where θ is the fraction indicating the attitude lenders are likely to take in extending short-term debt against the individual's holding of near-money liquid and nonliquid assets (including his future earning power). We shall take $0 \leq \theta \leq 1$.

Although L includes the adjustment made from the individual's own cash balance, we shall somewhat loosely refer to it as the individual's *maximum loan value* or his *line of credit.* In our discussion in this section, we shall take into account the effect of an individual's wealth on his optimum cash balance through its effect on his line of credit.

The borrowing or liquidation cost function $H(\cdot)$ will be assumed in this section to be applicable up to the point L. Beyond L, the nature of the penalty costs is assumed to change as expenditures have to be postponed or foregone or as profitable opportunities are lost. $Q(\cdot)$ will denote the loss (in utility or profits) involved in postponing or foregoing excess payments which cannot be met by liquidation of assets or by borrowing. In addition to the assumptions made in Section I regarding the penalty cost function $H(\cdot)$, we shall assume here that $Q''(\cdot) \geq 0$ and $Q'(\cdot) > H'(\cdot)$.

The individual's optimization problem may now be modified to be one of choosing a cash balance which maximizes the expected final wealth given by

$$E = W + S - \int_{M}^{L} H(x - M) f(x)\, dx - \int_{L}^{\infty} [H(L - M) + Q(x - L)] f(x)\, dx. \tag{10}$$

Or, again making the change of variable from x to $v = x + S$, we get

$$E = W + S - \int_{M+S}^{L+S} H(v - M - S) g(v)\, dv - \int_{L+S}^{\infty} [H(L - M) + Q(v - L - S)] g(v)\, dv. \tag{11}$$

Differentiating (11) partially with respect to M, keeping in mind that from (9) we have $\partial L/\partial M = 1 - \theta$, we obtain the expected marginal return from holding money in the present context:

$$\begin{aligned} E_M = {} & -r + (1 - r) \int_{M+S}^{L+S} H'(v - M - S) g(v)\, dv \\ & + \theta \int_{L+S}^{\infty} H'(L - M) g(v)\, dv \\ & + (1 - \theta - r) \int_{L+S}^{\infty} Q'(v - L - S) g(v)\, dv. \end{aligned} \tag{12}$$

The optimum cash balance of the individual is obtained again by setting $E_M = 0$.

The effect of a change in the individual's wealth on his money holding can be easily analyzed in the context of the present model by differentiating (12) partially with respect to W to obtain

$$\begin{aligned} E_{MW} = {} & -(1 - r) r \int_{M+S}^{L+S} H''(v - M - S) g(v)\, dv \\ & + \theta^2 \int_{L+S}^{\infty} H''(L - M) g(v)\, dv \\ & -(1 - \theta - r)(\theta + r) \int_{L+S}^{\infty} Q''(v - L - S) g(v)\, dv \\ & - (1 - r) r H'(0) g(M + S) \\ & -(1 - \theta - r)(\theta + r)[Q'(0) - H'(L - M)] g(L + S), \end{aligned} \tag{13}$$

where we have ignored the effect of wealth on expected payments which has been already discussed.

In equation (13), the second term is unambiguously nonnegative. The rest of the terms are either negative or of an indefinite sign. All the terms in (13) except the second will be negative if the quantity $(1 - \theta - r)$ is positive. Since an increase in the individual's money balance is most unlikely to *reduce* his total liquidity position, we shall assume that

$$1 - \theta - r > 0, \tag{14}$$

as this term precisely gives the effect of a change in the individual's money balance M on his total liquidity position, $(L + S)$.

Now comparing equation (13) with its counterpart in Section I, namely, equation (7), we see that the introduction of the constraint of the line of credit produces the possibility of a positive wealth effect through its effect on the line of credit. This would happen if in equation (13) the positive second term were to outweigh the rest of the (negative) terms.

Considering the second term in equation (13), it also becomes clear that an increase in the individual's wealth would increase his money demand by increasing his line of credit only if the marginal penalty cost were increasing and only if his line of credit were not so large as to become virtually irrelevant to his decision about optimal cash holding. If the individual's line of credit is so large that the probability of his current excess payments due exceeding it is zero or negligible, evidently the size of his optimum cash balance will not be influenced by changes in his line of credit. Even when the individual's line of credit is not so large, if the marginal penalty cost is constant, again his optimum cash balance will be independent of his line of credit.

The above discussion may be summarized by stating that the total net wealth of the individual will have a positive effect on his optimum cash balance by increasing his line of credit only if (1) the marginal penalty cost is increasing *and* (2) the individual's line of credit is small.

An increasing marginal penalty cost may be a realistic assumption if the loan market is not perfect in the sense that larger loans are more difficult to obtain (or if larger inadequacies of cash are thought of as being met by cashing less and less liquid assets, or by liquidating more and more lucrative investments). The influence of the individual's wealth constraint on his demand for money, originating in the imperfections of the loan market, has been pointed out by Chow (1966). However, our results go beyond the point made by him. He argued that if the capital market is perfect in the sense that loans can be extended at a given rate of interest (constant marginal penalty cost), then the demand for money will be determined by permanent income or human and nonhuman wealth of the individual. If, on the other hand, his borrowing cannot exceed his net nonhuman wealth (accumulated past savings), then money demand will be determined by his nonhuman wealth. That is, the imperfection in the capital market will simply determine which measure of wealth will determine the economic agent's demand for money. The result of our present section goes further than this statement. It says that neither total wealth nor nonhuman wealth will have a *positive* effect on the economic agent's money demand unless capital market imperfection is present. Furthermore, it must be emphasized that even this argument can explain the dependence of money demand on the wealth variable only for small wealth owners (whose line of credit may be expected to be small). For large wealth owners (or for wealth owners with large lines of credit), the explanation of the observed wealth elasticity of demand for money (e.g., in Brunner and Meltzer 1963; Meltzer 1963*a*, 1963*b*) must be sought elsewhere.

III

In this section, we shall explicitly introduce the constraint of the individual's net wealth on his excess payments due. Correspondingly, the dependence of the probability distribution of his excess payments due on the individual's present net wealth will also be explicitly introduced. We shall focus on the indirect effect of wealth on money demand arising through its effect on the probability distribution function of excess payments due.[5] To simplify our problem, we shall assume the marginal penalty cost to be constant $(= h)$. Also, we shall not introduce here the constraint of the line of credit.

The individual's optimization problem considered here, then, is to hold a money balance which maximizes the expected final wealth given by

$$E = W + S - \int_M^W h(x - M) f(x, W)\, dx, \tag{15}$$

which, after integrating the last term by parts, reduces to

$$E = W + S - h(W - M) + \int_M^W hF(x, W)\, dx. \tag{16}$$

On partially differentiating equation (16) with respect to M, using the result that $\partial x/\partial M = -r$ from the definition of x in (1), and making the change of variable from x to v, we obtain the expected marginal return from money holding:

$$E_M = -r + h(1 - r)[1 - G(M + S, W)], \tag{17}$$

and the optimum money balance is given by setting $E_M = 0$. The effect of wealth on the expected marginal return of money holding is given by

$$E_{MW} = h(1 - r)\{[P'(W) - r]G_1(M + S, W) - G_2(M + S, W)\}, \tag{18}$$

---

[5] See Chitre (1972) for a somewhat different model which explicitly introduces the constraint of planned expenditure on money balances and takes into account the effect of the total expenditure planned over the income period on the probability distribution of different time paths of the payments due *during* the income period.

where G_1 and G_2 are the partial derivatives of the cumulative distribution function $G(v, W)$ with respect to its first and second arguments, respectively. For density functions of continuous variables, of course, $G_1 > 0$. But the sign of $G_2(M + S, W)$ can be either positive or negative.[6]

Comparing equation (18) with equation (8) for the case of constant marginal penalty cost, we see that taking into account the constraint of the individual's net wealth on his excess payments due gives rise to the possibility of a positive (or negative) wealth effect in addition to the positive wealth effect arising through the effect of wealth on the mean level of the individual's excess payments due. However, since the second term in (18) is likely to be positive for individuals with large money holdings and negative for those with very small money holdings, we can say that for a positive wealth effect on money demand, *irrespective of the level of the individual's money holding*, once again the importance of the effect of net wealth on the individual's planned (or expected) level of expenditure cannot be undermined.

IV

In this section we take up the formulation of our problem in terms of a more general utility function of final wealth, involving diminishing marginal utility and therefore risk aversion. We shall examine here, in the context of an inventory theoretic formulation, the relevance to the wealth effect on the individual's demand for money of the changing risk aversion of the individual as his wealth increases.

We shall denote the individual's utility function by $U(\cdot)$, and it will be a function of the end-of-period wealth as in the simple model discussed by Arrow (1971). Unlike Arrow's discussion, however, our formulation will emphasize the uncertainty of payments due and anticipated receipts rather than the capital risk of the risky (nonmoney) assets, as the main determinant of the demand for money. A more general analysis, of course, must combine the two kinds of uncertainty and possibly also that of the future rate of inflation as a determinant of an economic agent's asset structure. This more general formulation is not attempted here.

We shall assume that the marginal utility of wealth is positive and diminishing:

$$U'(\cdot) > 0, \qquad U''(\cdot) < 0. \tag{19}$$

---

[6] We have

$$E(v) = \int_{-\infty}^{W+S} v \cdot g(v, W)\, dv = W + S - \int_{-\infty}^{W+S} G(v, W)\, dv$$

on integration by parts. Therefore, if $E(v)$ is assumed to be uninfluenced by the level of the individual's wealth, W, we obtain

$$\int_{-\infty}^{W+S} G_2(v, W)\, dv = 0.$$

But no restriction is implied on the sign of $G_2(M + S, W)$.

It is well known from the work of Pratt (1964) and Arrow (1971) that the quantity

$$R_A(\cdot) = -\frac{U''(\cdot)}{U'(\cdot)} \tag{20}$$

may be used as a measure of the absolute risk aversion of the individual. It indicates the extent of unwillingness of the individual to take a fair bet of a fixed absolute amount. As we should expect an individual to become more willing to take a fair bet involving a fixed absolute amount as his wealth increases, it is generally assumed that the absolute risk aversion of an individual, $-\{[U''(\cdot)]/[U'(\cdot)]\}$, diminishes as his final wealth increases. We shall make this assumption in what follows.

We shall use a constant marginal yield on earning assets and a constant marginal penalty cost in the present section. We shall also ignore here the effect of wealth on the individual's planned expenditure as well as on the distribution function of the individual's excess payments due.

The individual's optimization problem is now considered to be one of choosing a cash balance M which maximizes his expected *utility:*

$$\begin{aligned} E = &\int_{-\infty}^{M+S} U(W + S - v)g(v)\, dv \\ &+ \int_{M+S}^{\infty} U[W + S - v - h(v - M - S)]g(v)\, dv. \end{aligned} \tag{21}$$

The expected marginal utility of money holding is obtained simply by differentiating (21) partially with respect to M and is given by

$$\begin{aligned} E_M = &-r\int_{-\infty}^{M+S} U'(W + S - v)g(v)\, dv \\ &+ [h(1 - r) - r]\int_{M+S}^{\infty} U'[W + S - v - h(v - M - S)]g(v)\, dv. \end{aligned} \tag{22}$$

Differentiating (22) partially with respect to M, it is easily established that the assumption of diminishing marginal utility of wealth implies that $E_{MM} < 0$. That is, the expected marginal utility of money holding diminishes as the individual holds a larger money balance. Therefore, the optimum cash balance is obtained by setting $E_M = 0$. From equation (22) it is clear that we must have

$$h(1 - r) - r > 0, \tag{23}$$

or else the individual's optimum cash balance will be always equal to zero.

To analyze the wealth effect on money demand, we obtain, by differentiating E_M in (22) with respect to W,

$$\begin{aligned} E_{MW} = {} & -r(1+r)\int_{-\infty}^{M+S} U''(W+S-v)g(v)\,dv \\ & + [h(1-r)-r](1+r+rh) \\ & \int_{M+S}^{\infty} U''[W+S-v-h(v-M-S)]g(v)\,dv \\ & - h(1-r)rU'(0)g(M+S). \end{aligned} \tag{24}$$

Which way does the downward-sloping curve of the expected marginal utility of money holding shift if the individual's wealth increases? It can be shown by using mathematical reasoning similar to that employed by Arrow (1971, p. 119) that if the individual's absolute risk aversion diminishes with his wealth, then, in fact,

$$\begin{aligned} E_{MW} & < E_{MW} + h(1-r)rU'(0)g(M+S) \\ & \leq -(1+r)R_A[(W-M)(1+r)]E_M = 0.^7 \end{aligned} \tag{25}$$

That is, an increase in wealth lowers the expected marginal utility of the existing (optimum) money holding. An individual subject to diminishing absolute risk aversion would therefore respond to an increase in his wealth by reducing his money balance. Far from being a luxury, money would be an inferior asset (under the realistic assumption of decreasing absolute risk aversion) if the important effect of wealth on money demand through the effect of wealth on current planned expenditures is ignored. Whatever absolute amount of financial penalty there is for not holding an adequate cash balance, it weighs more heavily on the less wealthy.

It may be pointed out here that, in Arrow's discussion, the wealth effect on money demand is ambiguous if the assumption of decreasing absolute risk aversion is made, whereas our result is stronger and unambiguously negative. Arrow obtains an unambiguous result (with a wealth elasticity of money demand in fact exceeding unity) on the assumption that the *relative* risk aversion of the individual must be increasing, but his discussion makes it clear that the assumption of decreasing absolute

---

[7] Notice that with decreasing absolute risk aversion, for $x \leq M$ we have

$$-\frac{U''[W+r(W-M)-x]}{U'[W+r(W-M)-x]} \leq R_A[(W-M)(1+r)];$$

and for $x \geq M$ we have

$$-\frac{U''[W+r(W-M)-x-h(x-M)]}{U'[W+r(W-M)-x-h(x-M)]} \geq R_A[(W-M)(1+r)].$$

The result in equation (25) of the text follows immediately on making use of these inequalities.

risk aversion as well as that of increasing relative risk aversion must hold in realistic situations, since, as stated earlier, the assumption of increasing absolute risk aversion has the absurd implication that with an increase in wealth the individual becomes more unwilling to take a fair bet involving a fixed amount. Our result shows that if the indirect effect of wealth through its effect on planned expenditure is ignored, then for the wealth effect on money demand to be positive it is necessary that the individual's absolute risk aversion be increasing.

V

In our discussion so far, we have not considered the influence of the individual's wealth on his evaluation of the penalty costs of not having an adequate money balance. A fundamental sense in which the individual's wealth would influence the marginal penalty cost is the following. Suppose that not having an adequate money balance requires the individual to postpone expenditures in excess of his present money holding (including any receipts made by him). Then the marginal penalty cost is the excess of the marginal utility of present consumption over the (discounted) marginal utility of future consumption. But a higher level of wealth means that the *marginal* utility of future consumption is lower (because of diminishing marginal utility of consumption), and hence the marginal penalty cost is larger. Taking account of this influence of wealth on the marginal penalty cost does, of course, lead to a positive effect of wealth on the individual's money demand. Even in the context of this model the negative effect of wealth on money demand through the individual's expected receipts, exhibited in Section I, continues to be present, and the net effect is indeed indeterminate. What is more important, the introduction of near-money assets, which can enable the individual to avoid the postponement of expenditure by incurring the pecuniary penalty cost involved in borrowing against or in liquidating the near-money assets, makes a drastic difference to the conclusion. The wealth effect on money demand in this case is once again negative.

We shall examine these aspects of the wealth effect on money demand in the present section. We shall do this in the context of a simple two-period model.

As we shall want to distinguish between the utility of the present and that of the future consumption, we shall denote by $U(\cdot)$ the utility index function of one-period consumption expenditure. Future utility will be discounted by the discount factor α.

If the expected marginal yield on the nonmoney assets is denoted by r as before, the receipts during the current period will be given by $[r(W - M) + \eta]$ where η may be considered to be the stochastic disturbance term in the current receipts. The probability density function of η will be denoted by $g(\eta)$. Receipts may be negative as well as positive, since losses on *income* account, originating from the individual's real assets, are included in total receipts. The current-period consumption expenditure will be denoted by ξ and will be assumed to be a random variable with a probability density function $f(\xi)$.

Consider first the optimization problem where borrowing against or liquidating nonmoney assets to meet unforeseen expenditures is not possible and all expenditures beyond the individual's cash position (including receipts made by him) have to be postponed. The discounted expected utility which the individual seeks to maximize is then given by

$$\begin{aligned} E = {} & \int_{-\infty}^{\infty}\int_{0}^{M+Z+\eta} [U(\xi) + \alpha U(W + Z + \eta - \xi)] f\cdot g \, d\xi \, d\eta \\ & + \int_{-\infty}^{\infty}\int_{M+Z+\eta}^{\infty} [U(M + Z + \eta) + \alpha U(W - M)] f\cdot g \, d\xi \, d\eta, \end{aligned} \tag{26}$$

where we have let

$$Z = r(W - M) \tag{27}$$

for convenience.

In this case, it is easy to check that the effect of the individual's net wealth on his expected marginal utility of money holding is given by

$$\begin{aligned} E_{MW} = {} & \Bigg[\int_{-\infty}^{\infty}\int_{0}^{M+Z+\eta} - \alpha r(1 + r) U''(W + Z + \eta - \xi) f\cdot g \, d\xi \, d\eta \\ & - \int_{-\infty}^{\infty}\int_{M+Z+\eta}^{\infty} \alpha U''(W - M) f\cdot g \, d\xi \, d\eta \\ & + \int_{-\infty}^{\infty} \alpha r(1 - r) U'(W - M) f(M + Z + \eta) g \, d\eta \Bigg] \\ & + \Bigg[\int_{-\infty}^{\infty}\int_{M+Z+\eta}^{\infty} (1 - r) r U''(M + Z + \eta) f\cdot g \, d\xi \, d\eta \\ & - \int_{-\infty}^{\infty} (1 - r) r U'(M + Z + \eta) f(M + Z + \eta) g \, d\eta \Bigg]. \end{aligned} \tag{28}$$

The terms in the first square bracket give the positive wealth effect on money demand emphasized in the introduction to this section. As explained there, an increase in the individual's wealth increases the marginal penalty cost in the sense of the sacrifice of utility involved in postponing expenditures. To this extent, the expected marginal utility of money holding is increased due to an increase in the individual's wealth.

The terms in the second square bracket give the negative effect of wealth on money demand arising because an increase in wealth also increases the expected current receipts of the individual. This latter effect has already been exhibited in equation (7). Here we have its extension to the present model. The total wealth effect, of course, is indeterminate here.

Now consider a slightly different model in which the individual can avoid postponing expenditure by borrowing against or liquidating the nonmoney assets. For simplicity, we shall ignore here the constraint of the line of credit on such cash adjustment, which we have analyzed in Section II. We shall assume that the marginal borrowing cost is constant $(= h)$.

The individual's optimization problem now consists of maximizing the discounted expected utility given by

$$\begin{aligned} E = & \int_{-\infty}^{\infty} \int_{0}^{M+Z+\eta} [U(\xi) + \alpha U(W + Z + \eta - \xi)] f \cdot g \, d\xi \, d\eta \\ & + \int_{-\infty}^{\infty} \int_{M+Z+\eta}^{\infty} \{U(\xi) + \alpha U[W + Z + \eta - \xi \\ & \qquad - h(\xi - M - Z - \eta)]\} f \cdot g \, d\xi \, d\eta, \end{aligned} \tag{29}$$

where Z is given by (27) once again.

The expected marginal utility of money holding is given by

$$\begin{aligned} E_M = & \int_{-\infty}^{\infty} \int_{0}^{M+Z+\eta} -\alpha r U'(W + Z + \eta - \xi) f \cdot g \, d\eta \, d\xi \\ & + \int_{-\infty}^{\infty} \int_{M+Z+\eta}^{\infty} \alpha[h(1 - r) - r] \\ & \quad U'[W + Z + \eta - \xi - h(\xi - M - Z - \eta)] f \cdot g \, d\xi \, d\eta. \end{aligned} \tag{30}$$

The reader may observe at this stage the similarity of equation (30) with equation (22) in Section IV. By the reasoning given there, it would again follow that decreasing absolute risk aversion would lead to a *negative* wealth effect on money demand in the present case. Thus the presence of near-money assets allows the individual to avoid incurring one form of penalty cost, namely, the postponement of expenditure, and instead permits him to incur the penalty cost in the form of the higher cost of liquidating these assets. As his wealth increases, the individual with a decreasing absolute risk aversion finds himself more willing to bear this fixed financial penalty cost and hence willing to hold a smaller money balance.

References

Ando, A., and Modigliani, F. "The 'Life Cycle' Hypothesis of Savings: Aggregate Implications and Tests." *A.E.R.* 53 (March 1963): 54–84.

Arrow, K. J. *Essays in the Theory of Risk Bearing.* Chicago: Markham, 1971.

Arrow, K. J.; Karlin, S.; and Scarf, H. *Studies in the Mathematical Theory of Inventory and Production*. Stanford, Calif.: Stanford University Press, 1958.

Baumol, W. J. "The Transactions Demand for Cash: An Inventory Theoretic Approach." *Q.J.E.* 66 (November 1952): 545–56.

Brunner, K., and Meltzer, A. H. "Predicting Velocity: Implications for Theory and Policy." *J. Finance* (May 1963), pp. 319–59.

Chitre, V. S. "A Dynamic Programming Model of Demand for Money with a Planned Total Expenditure." *Internat. Econ. Rev.* 13, no. 2 (June 1972): 303–23.

Chow, G. C. "On the Long Run and Short Run Demand for Money." *J.P.E.* 74, no. 2 (April 1966): 111–31.

Feige, E. "Expectations and Adjustments in the Monetary Sector." *A.E.R.* 57 (May 1967): 462–73.

Friedman, M. "The Quantity Theory of Money—a Restatement." In *Studies in the Quantity Theory of Money*. Chicago: Univ. Chicago Press, 1956.

Meltzer, A. H. "The Demand for Money: The Evidence from the Time Series." *J.P.E.* 71, no. 3 (1963), 223–34. (*a*)

———. "The Demand for Money: A Cross Section Study of Business Firms." *Q.J.E.* 77 (August 1963): 405–32. (*b*)

Miller, M. H., and Orr, D. "A Model of the Demand for Money by Firms." *Q.J.E.* 80 (August 1966): 413–35.

Pratt, J. W. "Risk Aversion in the Small and in the Large." *Econometrica* 32 (1964): 122–36.

Teigen, R. L. "Demand Functions and Supply Functions for Money in the United States: Some Structural Estimates." *Econometrica* 32, no. 4 (October 1964): 476–509.

Tobin, J. "The Interest Elasticity of Transactions Demand for Cash." *Rev. Econ. Statis.* 38, no. 3 (August 1956): 241–47.

Tsiang, S. C. "The Precautionary Demand for Money: An Inventory Theoretical Analysis." *J.P.E.* 77 (February 1969): 99–117.

Turvey, R. *Interest Rates and Asset Prices*. London: Unwin, 1960.

Communications

On the General Equilibrium Analysis of Tax Incidence

J. Gregory Ballentine
Wayne State University

Ibrahim Eris
Universidade de Sao Paulo

I. Introduction

The virtual avalanche of recent literature on the effects of factor-market distortions has employed a variety of algebraic and geometric models.[1] By far the most prominent of such models in the public-finance literature has been the two-factor, two-good model presented by Arnold Harberger (1962).[2] His analysis has been modified and extended by many authors, but a crucial limitation of the model has been left unexplored.

This limitation of the Harberger model is that, due to the formulation of the demand equation, the model is valid only if all partial tax rates (or other price distortions), including the tax being analyzed, are zero. As a result, the model can only show the effects of introducing an *infinitesimally* small tax into a taxless world.

We explicitly insert nonzero taxes into Harberger's model and show that what is missing from that model is the effect on incidence of the tax-induced loss in real income. Though our model, like Harberger's, uses differential equations, we are able to derive theoretical results on the

---

The authors would like to thank Professors Charles McLure of Rice University and Wayne Thirsk of The University of Manitoba for their valuable comments. The authors accept responsibility for any errors remaining in the paper.

[1] See, for example, Johnson (1966), Johnson and Mieszkowski (1970), Herberg, Kemp, and Magee (1971), Jones (1971), and Magee (1971). Magee (1973) provides a survey of this literature.

[2] See McLure (1974) for a geometric exposition of Harberger's algebraic model.

[*Journal of Political Economy*, 1975, vol. 83, no. 3]
© 1975 by The University of Chicago. All rights reserved.

incidence and other effects of *finite* taxes. We explicitly review a few of the propositions developed in the Harberger literature to show how our model is more powerful in this respect. Further, we show how the zero tax requirement affects Harberger's estimates of the incidence of the corporation income tax.

II. The Model

The Harberger model requires two linearly homogeneous production functions for the goods X and Y, fixed endowments of the two factors K and L, perfectly competitive profit maximization, and consumer utility maximization. We shall write P_X, P_Y, and r for the prices of X, Y, and capital services (net of taxes) respectively. (We take labor to be the numeraire good.) For the moment we shall examine only one tax—a tax on capital in X, t_{KX}.[3]

The production equations of the model are identical with those presented by Harberger (1962, pp. 224–26). However, since he assumes $t_{KX} = 0$, the formula for capital's share (Θ_{KX} in our notation) which he uses is simply $r_{KX}/P_X X$. Because it is the producers' capital cost, and not just the net of tax payment to capitalists, which enters the profit-maximizing equations, the relevant formula for Θ_{KX} when $t_{KX} > 0$ is

$$r(1 + t_{KX})K_X/P_X X.^4$$

We use the latter formula, and thus Θ_{KX} denotes the gross of tax share of capital in X.

It is the demand equation of the model which changes when the tax assumption is dropped. Following Harberger, we assume that all individuals and the government consume identically. This allows us to write a single aggregate demand equation in output prices and disposable incomes. Thus

$$X = f(P_X, P_Y, rK + L + rt_{KX}K_X), \tag{1}$$

where $rK + L$ and $rt_{KX}K_X$ are private and government disposable incomes, respectively.[5]

---

[3] This slightly simplifies the algebra and allows us to emphasize that what is assumed by the Harberger model is not only that all other tax rates be zero, but that the very tax being analyzed must also be zero. In a final section we discuss differential tax incidence, and there we also consider a nonzero excise tax on X. Using our equations, which modify the Harberger model to accept nonzero taxes, the reader can easily substitute any combination of taxes which he feels is relevant for particular analyses.

[4] Harberger assumes all prices are equal to one, thus defining units of X, Y, K, and L. This allows him to consider the introduction of an infinitesimal ad valorem or per-unit tax on capital simultaneously, since in that case they are identical. We do not set prices equal to one and explicitly enter the tax as an ad valorem tax.

[5] For analytical simplicity and to maintain consistency with the Harberger model, we treat the private and public sector as if there is only one consumer in the economy. If

Differentiating (1) and separating into substitution and income effects, we obtain

$$X^* = \varepsilon(P_X^* - P_Y^*) + m/X[(K + t_{KX}K_X)\,dr + rK_X\,dt_{KX} + rt_{KX}\,dK_X$$
$$-X\,dp_X - Y\,dP_Y] \tag{2}$$

where an asterisk denotes percentage change, ε is the income-compensated elasticity of demand, and m is the marginal propensity to consume good X.

Using (2) and the production equations of the model, we solve for the basic-incidence equation

$$\eta_{r,t_{KX}} = \frac{r^*}{(1 + t_{KX})^*}$$
$$= \frac{\varepsilon\Theta_{KX}A + \sigma_X B - mP_X C\Theta_{KX}\sigma_X L_X/L_Y}{\varepsilon(\Theta_{KY} - \Theta_{KX})A - \sigma_X B - \sigma_Y + mP_X C\Theta_{KX}[\sigma_X(L_X/L_Y) + \sigma_Y]}, \tag{3}$$

where σ_X and σ_Y (defined as negative numbers) are the elasticities of substitution of capital for labor in X and Y, respectively; $A = (K_X/K_Y) - (L_X/L_Y)$; $B = \Theta_{LX}(K_X/K_Y) + \Theta_{KX}(L_X/L_Y)$; and $C = t_{KX}/(1 + t_{KX})$.

In the Harberger model the demand equation is written simply as[6]

$$X^* = \varepsilon(P_X^* - P_Y^*). \tag{4}$$

Using this demand equation and the production equations, the formula for $\eta_{r,t_{KX}}$, presented in the literature using the Harberger model, omits the last term in the numerator and denominator of (3). In what follows we shall call these two terms the *net-income terms*. It is the absence of any income effect in the demand equation of the Harberger model which

---

one assumes that marginal propensities to consume out of income for every consumer are the same, then there is no loss of generality as a result of the treatment given here. All one needs to do is to interpret the aggregate price elasticities appearing below as a weighted average of the individual ones. Should marginal propensities be different, the analysis would involve considerable additional complications. (See Mieszkowski 1967, p. 262).

[6] In the Harberger-model literature, Mieszkowski first explicitly identified ε in this demand equation as the income-compensated elasticity of demand (1967, p. 260, n. 10). However, Harberger's discussion of the relationship between ε and the elasticity of substitution in consumption makes it clear that he also intends ε to be the income-compensated elasticity of demand (1962, p. 228). While identifying ε, Mieszkowski also argues that if the government's and private sector's marginal propensities to consume X are the same, the income effects of government and private-sector demand just cancel out (1967, p. 260, n. 10). (Harberger did not explicitly show this, but his discussion of his demand equation also suggests that assuming equal marginal propensities to consume is sufficient to cancel out all income effects [1962, p. 224].) Of course, our assumption that the government and private demand functions are identical implies that the private and government marginal propensities to consume X are identical. In what follows we show that equal marginal propensities to consume is *not* a sufficient condition to cancel out all income effects, and that the additional condition $t_{KX} = 0$ is crucial.

crucially depends on the zero tax assumption *and* causes the omission of the net-income terms from (3).

To get a better idea of what the income effect represents and why it appears in the demand equation, we shall use the production equations of the model to simplify the income effect in the demand equation. Doing so, the income effect may be written as $(m/X)(rt_{KX}\, dK_X)$. The term rt_{KX} is simply the difference between the value of the marginal product of capital in industry X, $[r(1 + t_{KX})]$, and its value in industry Y, (r). Thus, the whole term in brackets shows the real-income loss as capital flows from industry X, where the value of its marginal product is higher, to industry Y, where it is lower.[7]

Along with the change in relative product prices operating through the Slutsky substitution term, the real-income loss changes the demand for X and Y and then, through the resulting changes in outputs, affects relative factor returns.[8] The net-income terms provide precisely this link between the loss of real income and changes in factor returns.[9] Obviously,

[7] The income effect as it appears in (2) is, of course, the Hicksian compensating variation (Hicks 1959, p. 99). It is an often-used measure of welfare change or the *excess burden of taxation*. In fact, this measure has been used along with the Harberger model to estimate the excess burden of the corporation income tax (Harberger 1966) and to discuss the relative excess burdens of tariffs and various taxes (Mieszkowski 1966). Thus, those studies attempt to measure real-income losses, while assuming that demand can be fully described by an income-compensated demand function, i.e., a demand function which assumes no real-income loss.

[8] Johnson and Mieszkowski present a similar discussion of the influence of the distortion-induced real-income loss on relative factor rewards in the context of a geometric model. They then use the Harberger model for an algebraic analysis and argue that the absence of such a real-income loss from that model is due to the use of differential equations (1970, pp. 541–47). It is not—it is solely the zero price-distortion assumption which causes the omission.

[9] Mieszkowski (1967) has argued that the effect of a partial factor tax on factor rewards can be divided into two separable effects—the output effect, which is the change in factor rewards due to the change in output of X and Y, and the substitution effect, which is the change in factor rewards as one factor is substituted for another at constant output. He associates the first and second terms of the numerator of (3) with these two effects, respectively. Krauss and Johnson (1972, p. 347) have argued that this separation is not particularly useful, partially because the output effect is not as clear-cut as Mieszkowski implies. The appearance of an additional term in our equation for $\eta_{r,t_{KX}}$ is indicative of the fact that Mieszkowski's characterization of the output effect is, indeed, oversimplified. He only takes into account the tendency to substitute Y for X in consumption (which is all that appears at the zero tax point). However, a part of the overall output change is the tendency to reduce consumption of both X and Y as real income falls due to the tax distortion (i.e., due to the excess burden of the tax). The additional term in our equation for (3) can be viewed as part of this overall-output effect. An important aspect of Krauss and Johnson's objection to Mieszkowski's characterization of the output effect is that his analysis does not admit of the possibility that output (and thus consumption) of *both* X and Y may fall. Their geometric analysis does indicate this possibility. They argue that this difference is at least partially due to Mieszkowski's use of differential equations. Once more, that is not the case; it is actually the zero tax assumption which causes this. It can be shown that our equations do admit of the possibility that output of both X and Y fall. This occurs when the effect of the real-income loss on consumption of Y is greater

when $t_{KX} = 0$, the value of the marginal product of capital is the same in both sectors, the income effect is zero, the net-income terms drop from (3), and, thus, the demand equation and the solution of the Harberger model apply.[10]

III. The Analysis of Finite Taxes

It is for the analysis of finite taxes that the difference between the formula for $\eta_{r.t_{KX}}$, which we present, and the one presented by those using the Harberger model, is crucial. Any analysis based on the Harberger model (i.e., using a derivative which must be evaluated at the point $t_{KX} = 0$) can only reach conclusions concerning the incidence or other effects of an *infinitesimal* tax. On the other hand, analysis using our equation (i.e., a formula for a derivative which is valid for any value of t_{KX}) can be used to describe the incidence of a *finite* tax.[11]

Theoretical propositions derived in the literature using the Harberger model have not always been acknowledged as having been proven for infinitesimal taxes alone. It would be useful to reexamine all of those propositions using equation (3) to determine which are valid for finite taxes and which are valid only for infinitesimal taxes. Clearly, we cannot do that here. However, we shall briefly consider the 10 propositions which

---

than the substitution of Y for X, due to the rise in P_X/P_Y. Further, this is only possible in the case of a partial factor tax which induces a factor-market distortion which, as is well-known, causes the transformation curve to shift in.

[10] That the income effect is zero when $t_{KX} = 0$, is simply the well-known proposition that real income is maximized when all price distortions are removed. It may be noted that if we had included a tax on capital in $Y(t_{KY})$, the income effect of the demand equation would be $r(t_{KX} - t_{KY})\, dK_X$. Thus, at any point at which $t_{KX} = t_{KY}$, the income effect would be zero, since the value of the marginal product of capital will be the same in both sectors. An overall finite income tax on capital can be expressed as two taxes, t_{KX} and t_{KY}, such that $t_{KX} = t_{KY}$. As is well known, in simple models in which the total supply of capital is fixed, this tax induces no excess burden. The income effect *throughout* the application of such a finite tax will be zero, and the Harberger model, without any modification, can analyze this and other finite *general* taxes.

[11] An intuitive nonmathematical justification for the ability of our model to analyze finite taxes, while the Harberger model cannot, relies on the fact that the effect of *any* finite tax can be divided into the sum of the effects of a series of infinitesimal taxes. The first of this series shows the effect of applying an infinitesimal tax when that tax is zero; the latter terms in the series show the effect of increasing an existing tax by infinitesimal amounts. The Harberger model solves for the first term in this series, and thus, for example, it can show that, under certain conditions, capital will bear more of the burden of an infinitesimal tax on capital in X than labor will (i.e., $\eta_{r.t_{KX}} < 0$ when $t_{KX} = 0$). All this means is that the first term in the series showing the effects of a finite t_{KX} on r is negative. Under the same circumstances, we may be able to show that our general formula for $\eta_{r.t_{KX}}$, valid for any value of t_{KX}, is negative. In such a case, we would have shown that capital will bear the burden of *increasing* any finite tax by an infinitesimal amount. This means that *each* term in the series discussed above is negative. From this we can conclude that the sum of the series is negative and, thus, that capital will bear the burden of any *finite* tax on capital in X. Clearly it is the zero tax assumption and not the use of derivatives which restricts the Harberger model.

Harberger proved in his 1962 paper (1962, pp. 227–30) in order to show how previous analyses can be generalized.[12]

Harberger's first proposition states that if X is the capital-intensive industry, then capital will bear more of the burden than labor of an *infinitesimal* tax on capital in industry X (i.e., $\eta_{r,t_{KX}}$ *at the point* $t_{KX} = 0$ is negative). Using our model, we can prove that when X is the capital-intensive industry, capital will bear more of the burden than labor of *any finite* tax on capital in industry X (i.e., this states that when $A < 0$, $\eta_{r,t_{KX}}$ as written in eq. [3] is negative). (Jones [1971] obtains this result also.)

Without proof, we shall only note that Harberger's propositions 3, 4, 5, and 8 can be similarly extended to apply for finite taxes, while 2, 6, and 7 can be proven *only* for infinitesimal taxes.

Proposition 9 is of particular interest, and we shall examine it more completely. That proposition states, in effect, that if both production functions are Cobb-Douglas and the elasticity of substitution of X for Y in consumption (σ_D) is -1 (thus $\sigma_X = \sigma_Y = \sigma_D = -1$), then the fall in the payments to capital induced by an infinitesimal tax on capital in X equals the rise in government revenue induced by that tax.[13] In fact, our equations show that this proposition can be extended to state that, under the above assumptions, the fall in the payments to capital due to any finite tax on capital in X exactly equals the government revenues from that tax.

To see this, let us first consider the values of $\eta_{r,t_{KX}}$ under the assumption $\sigma_X = \sigma_Y = \sigma_D = -1$. Since $\sigma_D = -1$ implies that the utility function underlying the demand equation is homothetic, one obtains $mP_X = P_X X/(P_X X + P_Y Y)$. Furthermore, as has often been demonstrated,[14] $\varepsilon = P_Y Y/(P_X X + P_Y Y)\sigma_D$. Substituting these into (3), we get $\eta_{r,t_{KX}} = -K_X/K$, and thus the fall in payments to capital is

$$\frac{dr}{dt_{KX}} K = \eta_{r,t_{KX}} \frac{rK}{1 + t_{KX}} = -\frac{rK_X}{1 + t_{KX}}.$$

---

[12] Before considering Harberger's propositions, one additional difference between the solution of his equations and ours should be noted. In both solutions the term $(\Theta_{KY} - \Theta_{KX}) A$ appears. In Harberger's solution, since $t_{KX} = 0$, this term is necessarily negative. However, as has been pointed out by Jones (1971), Magee (1971), and others, when $(t_{KX} \neq 0)$, this term may be positive. Such a case implies that relative factor intensities in value terms differ from relative factor intensities in physical terms. We shall assume that such a "value reversal" does not occur.

[13] Harberger states the conclusion of this theorem as "capital will bear precisely the full burden of the tax" (1962, p. 230). Since, by considering only the zero tax point, "excess burden" (i.e., the welfare loss due to the distortion) is omitted, the "full burden of the tax" borne by the private sector is simply the transfer to the government (dI_G). However, since we consider nonzero tax points where real-income loss (excess burden) does appear in the derivatives, we state the theorem simply in terms of the equivalence of the change in factor payments to capital and the change in government revenues.

[14] See, for example, Harberger (1957, p. 514).

Now turning to dI_G/dt_{KX}, where I_G is government revenues, we have

$$\begin{aligned}\frac{dI_G}{dt_{KX}} &= rK_X + t_{KX}K_X\frac{dr}{dt_{KX}} + t_{KX}\frac{r\,dK_X}{dt_{KX}} \\ &= rK_X + \eta_{r,t_{KX}}\frac{rt_{KX}K_X}{1+t_{KX}}(1 + B_1 + B_2) + \frac{rt_{KX}K_X}{1+t_{KX}}B_2,\end{aligned} \tag{5}$$

where $B_1 = -(\sigma_Y/A)$ and $B_2 = -(L_X\sigma_X/L_YA)$. Utilizing the just-derived value for $\eta_{r,t_{KX}}$ and substituting $\sigma_X = \sigma_Y = -1$, we obtain also $dI_G/dt_{KX} = rK_X/(1 + t_{KX})$, which completes the proof.

Harberger's proposition 10 is a generalization of proposition 9, stating that the conclusion of 9 holds for any values of σ_D, σ_X, and σ_Y as long as $\sigma_D = \sigma_X = \sigma_Y$. As can be seen from equation (3) above, the value of $\eta_{r,t_{KX}}$ is invariant with respect to proportionate changes in σ_D, σ_X, and σ_Y. The value of dI_G/dt_{KX}, however, as shown in (5), is not invariant with respect to proportionate changes in σ_D, σ_X, and σ_Y, except when $t_{KX} = 0$. Thus, proposition 10 can be proven only for an infinitesimal tax case.

IV. A Reestimation of the Incidence of the Corporation Income Tax

Probably the most important empirical application of Harberger's model has been his own use of it to estimate the shifting of the corporation income tax (1962, pp. 230–35). As a result of his analysis, he concluded that that tax is fully borne by capital. However, Harberger's analysis is based on a misuse of his model, which is directly related to the zero tax requirement.

Both Harberger's model and the one presented here provide only equations for the slope of the relationship between, say, r and t_{KX}. If t_{KX} is "small," then an empirical estimate of its incidence may use only one such equation—for example, pretax or posttax values (though it is incumbent on the analyst to justify his criteria for determining the "smallness" of a tax). On the other hand, if the tax is "large," then the estimate of the slope at one point may not provide a good approximation for the incidence of that tax. In such a case, it is useful to provide estimates based on, at least, pretax and posttax values. Harberger raised this same issue, stating that "if the corporation income tax were of small magnitude the pre-tax values of [the variables] . . . would all be very close to their post-tax values . . . However, the tax is in fact substantial in the United States. I have accordingly decided to use two alternative sets of values [pre-tax and post-tax]" (1962, p. 231). He then proceeds to insert the posttax values into the equation derived from his model, which, as we have seen, is invalid when there are taxes.

To indicate the empirical relevance of the appearance of the net-income terms in (3), it would be useful to compare Harberger's posttax estimates, which ignore those terms, with estimates based on equation (3). However, such an approach would not isolate the influence of those terms, since (3) differs from the equation Harberger uses in one other respect. For his posttax estimates Harberger calculates Θ_{KX} as net of taxes. The term Θ_{KX} enters the model through the profit-maximizing conditions where it must reflect producers' capital costs in X, $[r(1 + t_{KX})\ K_X]$, not simply the returns to capitalists (rKx). Thus, Θ_{KX} must properly be calculated as gross of taxes. A simple comparison of Harberger's estimates with those based on (3) would combine the effects of the net-income terms and conversion of Θ_{KX} from net to gross of taxes. Therefore, we shall redo Harberger's estimates, only converting Θ_{KX} to gross of taxes, and compare those values with values obtained using (3).

Letting X be the corporate sector and Y the noncorporate, and using Harberger's 1953–55 data, in billions of dollars we have $rK_X = 20$, $rt_{KX}K_X = 20$, $wL_X = 200$, $rK_Y = 20$, and $wL_Y = 20$. From these one can calculate every variable in (3) except σ_X, σ_Y, ε and mP_X. Harberger considers different values for σ_X, σ_Y, and σ_D. From σ_D he estimates ε by the formula presented in the previous section. Since mP_X does not appear in his equations, we must provide a separate estimate for it. However, Harberger's treatment of σ_D as a constant for before- and after-tax values is consistent with a CES utility function which enables us to use the formula $mP_X = P_XX/(P_YY + P_XX)$.

Beyond calculating $\eta_{r,t_{KX}}$, Harberger provides calculations of $-K\ dr/dI_G$, that is, the fall in factor payments to capital as a proportion of the increase in government revenues.[15] However, Harberger calculates dI_G, even when $t_{KX} = 1$, as simply $dI_G/dt_{KX} = rK_X$, that is, as if $t_{KX} = 0$ (see eq. [5]). Thus the ratio he calculates is $(K/rK_X)(dr/dt_{KX})$.

In table 1 the first column shows $\eta_{r,t_{KX}}$ calculated omitting the net-income terms (i.e., correcting Harberger's calculations by converting Θ_{KX} to gross of taxes). The second column shows estimates of $-K\ dr/dI_G$ in percentage terms using column 1 values of $\eta_{r,t_{KX}}$ and Harberger's zero tax formula for determining that ratio. The third column provides the correct values of $\eta_{r,t_{KX}}$, and the fourth, the corresponding value of $-K\ dr/dI_G$ using (5).

The most striking result of this comparison is that, for this set of data, omission of the net-income terms results in 16–20 percent overestimation of $\eta_{r,t_{KX}}$ (in absolute value). Further, such omission and calculation of $-K\ dr/dI_G$ on the basis of $t_{KX} = 0$ when, actually, $t_{KX} = 1$, also result in large errors.[16]

---

[15] As in proposition 9 discussed in Section III above, he terms this "capital's share of the burden." However, when $t_{KX} \neq 0$, dI_G is not the full burden borne by the private sector, since then there is an "excess burden."

[16] It should be emphasized that this direction of bias is not a general result. With different data, the omission of the net-income terms can cause a bias in just the opposite direction. Because the net-income terms enter the numerator and denominator of (3),

TABLE 1

A CORRECTION OF HARBERGER'S INCIDENCE ESTIMATES

	$\eta_{r,t_{KX}}$*	$-K\,dr/dI_G$†	$\eta_{r,t_{KX}}$	$-K\,dr/dI_G$
1. $\sigma_X = \sigma_Y = \sigma_D = -1$	−.58	116%	−.50	100%
2. $\sigma_X = \sigma_D = -1; \sigma_Y = -.5$	−.67	134	−.59	117
3. $\sigma_X = \sigma_Y = -1; \sigma_D = -.5$	−.64	128	−.57	111
4. $\sigma_X = \sigma_Y = -1; \sigma_D = -1.5$	−.53	106	−.44	88
5. $\sigma_Y = \sigma_D = -1; \sigma_X = -1.2$	−.63	126	−.55	121
6. $\sigma_Y = \sigma_D = -1; \sigma_Y = -.8$	−.52	104	−.44	78
7. $\sigma_Y = \sigma_X = -.66; \sigma_D = -1$	−.53	106	−.44	74

* Omitting the net-income terms.
† As Herberger calculated dI_G.

In the appendix to his paper, Harberger considers the possibility that the corporate sector charges a "monopoly markup" (1962, pp. 239–40). For this case he assumes that $P_X X = [r(1 + t_{KX})K_X + wL_X](1 + M)$, where M is the monopoly markup. Such a markup induces a product-market distortion in addition to, and independent of, the factor-market distortion induced by t_{KX}. As a result, even when $t_{KX} = 0$, the demand equation must show an income effect and net-income terms will enter $\eta_{r,t_{KX}}$.[17] Harberger again used a formula which omitted the net-income terms. The correct formula is

$$\eta_{r,t_{KX}} = \frac{\begin{array}{r}\varepsilon A(1 + M)\Theta_{KX} + \sigma_X B(1 + M) \\ - mP_X C\Theta_{KX}(L_X/L_Y)\sigma_X - mP_X M\sigma_X B\end{array}}{\begin{array}{r}\varepsilon A[\Theta_{KY} - \Theta_{KX}(1 + M)] - \sigma_X B(1 + M) - \sigma_Y \\ + mP_X C\Theta_{KX}[(L_X/L_Y)\sigma_X + \sigma_Y] + mP_X M(\sigma_X B + \sigma_Y)\end{array}}.$$

Taking one of Harberger's suggested values of M at 0.2, we found that the values of $\eta_{r,t_{KX}}$ do not change markedly from those derived in the nonmonopoly markup case. However, to the extent that capitalists receive not only the normal returns to capitalists, but also the monopoly profits,

---

and, more important, because the aspect of the overall output effect attributable to the net-income terms does not operate in the usual way of a substitution of Y for X, which then affects factor returns depending on factor intensities, but instead affects factor returns through a tendency to reduce consumption of both X and Y as real income falls, we have not developed a simple, intuitive rationale for the directional bias in this case. In this respect, however, we note that the *sign* of the difference between our general formula for $\eta_{r,t_{KX}}$ and Harberger's formula does *not* depend on the value of m or of t_{KX}. (This applies only for a partial factor tax. For a partial excise tax on X, which, of course, also induces some excess burden, a larger value of m increases the fall in consumption and output of X *along the unchanged transformation curve* and, therefore, increases in absolute value the change in r. The sign of the change in r depends only on relative factor intensities. Thus we can say that the zero tax formula for the incidence of a partial excise tax will always underestimate the fall or rise in r. On this point, see the formula for η_{r,t_X} given in eq. [7] and that given by Mieszkowski [1967, p. 254].)

[17] The income effect in the demand equation for this case is $(rt_{KX}\,dK_X + M\,dX)$. Thus, even when $t_{KX} = 0$, this term does not drop out.

the fall in total payments to capital as a percentage of the rise in government revenues will be significantly higher when a monopoly markup is present. This is because in the nonmarkup case, payments to capitalists fall only as r falls, while in the markup case they fall both as r falls and as monopoly profits fall. (With M, fixed monopoly profits fall if P_XX falls.) A more complete analysis of this situation would place the tax not only on rK_X, but also on monopoly profits, in which case the analysis concerning the fall in total profits as a percentage of tax revenues would be different.

V. Differential-Incidence Analysis

In the previous sections, we have discussed only balanced-budget incidence. That is, as a tax is imposed, government revenues increase and the government spends all of its revenues. The Harberger model has also been used to discuss differential tax incidence (see, for example, Mieszkowski 1966, 1967; Krauss 1972). Beyond the omission of the net-income terms, the zero tax assumption made in that analysis is particularly restrictive.

The additional restriction arises because, while differential-incidence analysis considers the substitution of one tax for another holding the government budget unchanged, the zero tax assumption requires that the government budget be zero. To develop a more general differential-incidence analysis, we shall explicitly consider, besides t_{KX}, an excise tax on X, t_X.

To substitute t_X for t_{KX} while holding the government budget unchanged in real terms, we require that the government revenues always be sufficient to purchase a given bundle denoted by $\bar{X}_G$, $\bar{Y}_G$.[18] Thus $rt_{KX}K_X + P_Xt_XX = P_X(1 + t_X)\bar{X}_G + P_Y\bar{Y}_G$. Implicitly differentiating this, we obtain

$$t_{KX}K_X\,dr + rK_X\,dt_{KX} + rt_{KX}\,dK_X + P_Xt_X\,dX + P_XX\,dt_X + t_XX\,dP_X = P_X\bar{X}_G\,dt_X + (1 + t_X)\bar{X}_G\,dP_X + \bar{Y}_G\,dP_Y. \quad (6)$$

Collecting terms and solving this, one can derive the rate at which t_X can be substituted for t_{KX} so as to leave the government budget unchanged:

$$\frac{(1 + t_{KX})^*}{(1 + t_X)^*} = -\frac{F\eta_{r,t_X} + P_X(X - \bar{X}_G)(1 + t_X)}{F\eta_{r,t_{KX}} + F + rK_X + (P_XX/A)(D_1 + D_2) + P_Y\bar{Y}_G\Theta_{KY}} \quad (7)$$

---

[18] By this we are not requiring the government to actually purchase $\bar{X}_G$ and $\bar{Y}_G$, only that it be just able to do so.

where

$$F = \Theta_{LX} r t_{XK} K_X + P_Y \bar{Y}_G(\Theta_{KX} - \Theta_{KY}) - (P_X X/A)(D_1 + D_2),$$

$$D_1 = \frac{t_{KX}}{1 + t_{KX}} \Theta_{KX} \sigma_X (L_X/L_Y) + t_X \sigma_Y B,$$

$$D_2 = \frac{t_{KX}}{1 + t_{KX}} \Theta_{KX} \sigma_Y + t_X \sigma_Y,$$

$$\eta_{r,t_{KX}} = \frac{\varepsilon A + \sigma_Y B - m P_X D_1}{\varepsilon(\Theta_{KX} - \Theta_{KY}) - \sigma_X B - \sigma_Y + m P_X (D_1 + D_2)},$$

and

$$\eta_{r,t_X} = \frac{\varepsilon A}{(\Theta_{KX} - \Theta_{KY})A - \sigma_X B - \sigma_Y + m P_X (D_1 + D_2)}.$$

Writing the numerator of equation (7) as α_1 and the denominator as α_2, we can express the differential incidence of t_{KX} and $t_X(\eta^D_{r,t_{KX},t_X})$ as $\eta^D_{r,t_{KX},t_X} = \eta_{r,t_X} - \eta_{r,t_{KX}}(\alpha_1/\alpha_2)$.

Differential incidence analyses in the Harberger literature (principally Mieszkowski 1967; Krauss 1972) have employed the zero tax assumption to reduce equation (6) to

$$rK_X\, dt_{KX} + P_X X\, dt_X = 0. \tag{8}$$

Obviously, this greatly simplifies the algebra, but it does so at considerable expense in terms of exposing the complexity of the differential-incidence problem.[19] Omitted from equation (8) and analysis based on it, but indicated in equation (6), are such considerations as the fact that (1) as t_X is raised, the price the government must pay for X will rise; (2) independent of the change in the excise tax, revenues from that tax will change as P_X and X change, due to the change in t_{KX}; and (3) similarly, independent of the change in t_{KX}, revenues from that tax will change as the change in t_X causes X and thus K_X and r to change, etc.

Thus for analysis of differential tax incidence, the zero tax requirement of the Harberger model is doubly restrictive. It causes the omission of the influence of real-income changes on tax incidence and also the omission of many responses in the government-budget equation, as two tax rates are varied.

References

Harberger, A. C. "Some Evidence on the International Price Mechanism." *J.P.E.* 65, no. 6 (December 1957): 506–21.

———. "The Incidence of the Corporation Income Tax." *J.P.E.* 70, no. 3 (June 1962): 215–40.

---

[19] For this zero tax case, as can be seen from eq. (6) or (7), α_1/α_2 becomes simply $1/\Theta_{KX}$.

———. "Efficiency Effects of Taxes on Income from Capital." In *Effects of Corporation Income Tax*, edited by Marion Krzyzaniak. Detroit: Wayne State Univ. Press, 1966.

Herberg, Horst; Kemp, M. C.; and Magee, Stephen P. "Factor Market Distortions, the Reversal of Relative Factor Intensities, and the Relation between Product Prices and the Equilibrium Outputs." *Econ. Rec.* 97 (December 1971): 518–30.

Hicks, John R. *A Revision of Demand Theory*. Oxford: Clarendon, 1959.

Johnson, H. G. "Factor Market Distortions and the Shape of the Transformation Curve." *Econometrica* 34 (July 1966): 686–98.

Johnson, H. G., and Mieszkowski, P. "The Effects of Unionization on the Distribution of Income: A General Equilibrium Approach." *Q.J.E.* 84 (November 1970): 539–61.

Jones, R. W. "Distortions in Factor Markets and the General Equilibrium Model of Production." *J.P.E.* 79, no. 3 (May/June 1971): 437–59.

Krauss, M. "Differential Tax Incidence: Large versus Small Tax Changes." *J.P.E.* 80, no. 3 (January/February 1972): 193–97.

Krauss, M., and Johnson, H. G. "The Theory of Tax Incidence: A Diagrammatic Analysis." *Economica* 39 (November 1972): 357–82.

McLure, C. E., Jr. "A Diagrammatic Exposition of the Harberger Model." *J.P.E.* 82, no. 1 (January/February 1974): 56–82.

Magee, S. P. "Factor Market Distortions, Production, Distribution and the Pure Theory of International Trade." *Q.J.E.* 85 (November 1971): 623–43.

———. "Factor Market Distortions, Production and Trade: A Survey." *Oxford Econ. Papers* 25 (March 1973): 1–43.

Mieszkowski, P. M. "The Comparative Efficiency of Tariffs and Other Tax-Subsidy Schemes as a Means of Obtaining Revenue or Protecting Domestic Production." *J.P.E.* 74, no. 6 (December 1966): 587–99.

———. "On the Theory of Tax Incidence." *J.P.E.* 75, no. 3 (June 1967): 250–62.

Peak-Load Pricing under Regulatory Constraint: A Proof of Inefficiency

Leonard Waverman

University of Toronto

There exists a lengthy literature discussing the regulated utility and the distortions in the capital-labor choice engendered by rate-of-return regulation (Bailey and Malone 1970; Baumol and Klevorick 1970; Atkinson and Waverman 1973). The literature indicates an incentive for the utility to overcapitalize, that is, use a capital-labor ratio at each level of output higher than that dictated by cost minimization. All of these analyses utilize a single-period model.

In a recent contribution to this *Journal*, Elizabeth E. Bailey utilizes an *n*-period model to examine the peak-load pricing policies of a utility regulated as to the maximum return obtainable on its rate base. Two very strong statements appear: "Although it is found that the firm increases its rate base by increasing capacity, there is no implication in the peak-load pricing models that the firm has an incentive to operate at a nonminimum-cost technology" (Bailey 1972, p. 677), and "The method of regulation thus provides an advantage in that it dampens the monopolistic tendency to restrict output, but it also entails the disadvantage that off-peak users do not share in the price reductions" (Bailey 1972, p. 678).

The impression is left with the reader that as soon as one extends the single-period model to *any* multiperiod peak-load pricing model, the inefficiency of rate-base regulation disappears. This statement, if true, would suggest that the earlier literature on rate-base regulation and its distortions is incorrect. It is the purpose of this comment to show that Bailey has implicitly used a production function entailing fixed proportions in both the short and the long run. A more general model will lead to the expected capital bias. If the firm in question has possibilities for ex ante substitutability in its capital-labor choice in the long run but

This comment was written while I was at the Public Economics Workshop at the University of Essex, July 1973. I would like to thank all participants for listening to these ideas, Tony Atkinson for helping to make them operational, Mel Fuss for smoothing out some rough spots, and David McNicol for pointing out an error in an original draft. All remaining errors are my sole responsibility.
[*Journal of Political Economy*, 1975, vol. 83, no. 3]
© 1975 by The University of Chicago. All rights reserved.

operates with fixed proportions in the short run (a putty-clay model), distortion of the kind discussed in the earlier literature exists in a multiperiod peak-load pricing model. Moreover, incorporating long-run ex ante substitutability between capital and labor leads to the conclusion that rate-of-return regulation benefits off-peak users as well as peak users.[1]

Bailey's Model

In the Bailey paper, the firm is hypothesized as maximizing profits subject to a regulatory constraint. Profits (¶) is the sum of revenue (R) less costs. Costs take two categories, operating expenses in each period $[C(q_1, \ldots, q_n)]$ and capacity costs (k), which depend only on the peak capacity (x):

$$\P = R(q_1, \ldots, q_n) - C(q_1, \ldots, q_n) - kx. \qquad (1)$$

I suggest that Bailey's conclusion that the multiperiod regulated utility will not inefficiently substitute capital for labor has been constructed in objective function (1).

The firm *cannot* substitute capital for labor. The cost function is separable into fixed costs and operating costs. Fixed costs depending only on peak capacity are not affected by output. Operating costs are not, however, functions of capacity (ultimate or utilized) but depend only on output in each period.[2] While this may be true ex post in any short-run period, the firm's ex ante decision of the optimal level of capital will generally depend on the relative prices of inputs and the effects of regulation.

What would the objective function (1) in Bailey's paper imply for the long-run production function in any single specific peak period i? The firm is using the available capital stock and hiring labor (operating expenses) in period i (Li) from competitive markets at a per unit price w: $q_i = L_i/a$, where a is the constant production coefficient in period i. Indeed, what Bailey is assuming is a fixed-proportions production function in the long run. Where the production function is one of fixed proportions, the inputs are perfect complements. The firm cannot substitute one input for the other. In this case, where both input prices are positive, changes in input prices do not affect the relative quantities of labor and capital hired. The effects of rate-of-return regulation are comparable to changes in the relative price of inputs. It is therefore not surprising that Bailey is able to prove that rate-of-return regulation does not lead to the substitution of capital for labor.[3]

---

[1] An original version of this paper was in error on this point. David McNicol pointed that error out in private correspondence with me and suggested the proof which follows eq. (17).

[2] This assumption of an additively separable cost function has a long history in the marginal-cost-pricing literature. Williamson (1966), for example, has the firm operating with constant operating costs of b per output unit and β per unit of capacity. Littlechild (1970) assumes constant operating costs of c_j in each jth period and constant unit capacity costs of β.

Is it reasonable to assume that fixed proportions in both the short and the long run typify the technological choices facing all public utilities? For some public utilities, for example, telephone systems, it is likely to be a good approximation, but only in the short run. At any point in time, capital is fixed and the output is a function of the number of operators brought in on a shift. For utilities such as electricity generation, even this short-run model is not apt. Each specific plant brought on stream has its own associated operating expenses. The plant used at the peak—gas turbines—is characterized by lower capital costs and higher operating costs than the base load equipment. To adequately represent the behavior of a regulated utility facing both peak and off-peak demands requires a less restrictive model than one which implicitly assumes fixed proportions.

A General Model

Instead of incorporating only a short-run decision model of the firm to examine peak-load pricing, a more realistic long-run model incorporating ex ante and ex post decisions would be desirable.

Let us make the assumption that capital is of a putty-clay nature. The firm chooses capital-labor ratios freely ex ante, but once decisions are made and capital installed, the capital-labor ratio is fixed ex post.

Denote capital by K, its unit purchase price (cost of capital, assuming equipment price is unity) from competitive markets by r, and denote labor by L, which is purchased from competitive markets at a price per unit of w. Output is Q. Periods are denoted by subscripts. Revenue, R, is the product of output, Q, and price, p, where p comes from an inverse demand function $p = p(Q)$.

The capital-labor ratio (K/L) for simplicity will be called k. With constant returns to scale, output per man (Q/L) is a function of k:

$$Q/L = f(k). \tag{2}$$

As was suggested above, Bailey implicitly assumed a fixed-proportions production function. In the following discussion, I will make a strong assumption to put this more general model within the confines of Bailey's paper. Assume that the firm chooses the capital-labor ratio ex ante so as to maximize profits and is able to vary capital and labor freely. Once a specific optimal capital-labor ratio is chosen, the firm operates in any single period as if it had a fixed-proportions production function, and the ratio of labor hired (L_t) to utilized capital (K_t) is a constant. Actually, this proportion (k) is itself a choice variable in the long-run ex ante case. If inefficiency can be demonstrated in the capital-labor choice for this specific model, an analogue to Bailey's model, the general result of efficiency in peak-load pricing models of regulated utilities cannot hold.

---

[3] McNicol (1973) has shown that a utility faced with fixed technological proportions in all periods can meet the regulatory constraint only by holding excess capital in all periods. The fixed-proportions model then is not essentially useful for analyzing the peak problems, for it does not allow a peak to exist.

The Unregulated Profit-Maximizing Firm

A profit-maximizing monopoly would set the following objective:[4]

$$\max \P = \sum_t R_t(Q_t) - \sum_t wL_t - rK + \sum_t v_t(K - K_t), \qquad (3)$$

where $\P$ is the present value of profits and discounting is ignored for simplicity.

The firm would maximize revenue less the costs of labor hired in each period less the cost of the capacity capital stock subject to the constraint that K_t, the capital used in any period, be no greater than allowed by capacity (K). The firm's choice variables, ex ante, are output in each period Q_t, capacity capital stock K, and the technique of production represented by the capital-labor ratio k. The assumptions are that capital and labor are freely substitutable ex ante within the limits of a production function; once a choice has been made, the firm uses capital and hires labor in any period so that the ratio of utilized capital to labor services is a constant (k).

Equation (3) can be rewritten substituting in k wherever possible, leading to

$$\max \P = \sum R_t - \frac{\sum wQ_t}{f(k)} - rK + \sum v_t\left(K - \frac{kQ_t}{f(k)}\right). \qquad (4)$$

Taking first derivatives with respect to the choice variables Q_t, K, and k and the shadow price v_t yields

$$\frac{\delta \P}{\delta Q_t} = MR_t - \frac{w}{f(k)} - v_t \frac{k}{f(k)} = 0, \qquad (5)$$

$$\frac{\delta \P}{\delta K} = -r + \sum_t v_t \leq 0, \qquad (6)$$

$$\frac{\delta \P}{\delta k} = \frac{wQ_t}{f^2} f_k - \frac{v_t Q_t}{f^2}(f - kf_k) = 0, \qquad (7)$$

---

[4] I assume that R and $\P$ have all the necessary properties of continuity, concavity, etc.

(where MR_i is the marginal revenue of output and f_k is the first derivative of Q/L with respect to k),

$$\frac{\delta \P}{\delta v_i} = \frac{kQ_i}{f(k)} - K \leq 0 \tag{8}$$

and its complementary condition that $v_i \geq 0$ if $K - kQ_i/f(k) \geq 0$ (the shadow price of capital is zero if capital is not fully utilized in the period). Examining (5) and (6), we find conditions identical to those reported in Bailey for the unregulated profit-maximizing case.

The price equations are obtained by summing over all the i periods in (5):

$$\sum_i MR_i = \sum_i w\frac{L_i}{Q_i} + \sum_i v_i\frac{K}{Q_i}. \tag{5a}$$

Marginal revenue in each period is set equal to the sum of marginal operating cost times the labor-output ratio plus the shadow value of capacity times the capital-output ratio. These weights on unit prices of labor and capital are due to the putty-clay assumption of choosing the technique of production ex ante. From (8) we see that the shadow value of capacity is nonzero if and only if capacity is being fully used in each period. Therefore, defining a peak period (and denoting it by subscript p) as one where capacity is fully used and an off-peak period as one where capacity is not fully utilized (and denoting it by subscript j) yields

peak period:

$$MR_p = w\frac{L_p}{Q_p} + v_p\frac{K}{Q_p}; \tag{5b}$$

off-peak period:

$$MR_j = w\frac{L_j}{Q_j}. \tag{5c}$$

Equations (5b) and (5c) constitute the familiar rule that only peak users pay for capacity.[5]

Equation (6) ensures that the sum of shadow values of capital equals the price of capital.

Equation (7) is the rule for efficiency; f_k is equal to the marginal product of capital. For the constant returns-to-scale production function, $f - kf$, $[Q/L - (K/L)f_k]$ must equal the marginal product of labor (denoted f_L), since, by Euler's theorem, $Lf_L + Kf_k = Q$. From (7),

$$wQ_if_k = v_iQ_i(f - kf_k). \tag{7a}$$

---

[5] Note that if the peak is characterized by different and more expensive labor, then the operating costs at the peak may be higher than off peak. However, if off-peak labor is overtime or nighttime labor, operating costs may be higher off peak. In general, w, the operating costs, could vary between periods. The essential results of this paper would, however, remain the same but even more notation would be required.

Summing over all i periods and rearranging terms yield

$$\sum_{i} \frac{f - kf_k}{f_k} = \frac{\sum_{i} wQ_i}{\sum_{i} v_i Q_i} = \frac{f_L}{f_k}. \tag{7b}$$

The sum $\sum_i v_i Q_i$ can be replaced by $r \sum_p Q_p$. The shadow price of capacity is nonzero ($v_i > 0$) only if capacity is fully utilized. Therefore, for off-peak periods $\sum_j v_j Q_j \equiv 0$. For peak periods, $Q_p = K$ (and capacity is constant). Therefore, $\sum_i v_i Q_i$ reduces to $\sum_p v_p Q_p$, which is equal to $r \sum_p Q_p$ from (6).

Equation (7b) reduces to

$$\frac{\sum Q_p}{\sum Q_i} \frac{f_L}{f_k} = \frac{w}{r}. \tag{9}$$

On average, for the unregulated monopoly the ratio of weighted marginal products of labor to capital is set equal to the ratio of relative prices. The weight is the proportion of time which is peak period (where capacity is fully utilized).

The firm must, before the set of periods, peak and off peak, choose a technique of production (k) which maximizes its profits. In so doing it adjusts the marginal products of the two factors by the average time where capital is fully utilized. Capital must be paid for each period. Labor is a function of output in each period. Therefore, the equality of the ratio of marginal products to respective input prices is adjusted for the percentage of time when capacity can be charged for. If all periods are peak, (9) reduces to the standard marginal condition.

The Regulated Utility

Assume that the profit-maximizing firm described in the previous section is regulated as to its maximum rate of return. This utility, able to earn at most s on its capacity, has the following maximization problem (where $\sum_i R_i - \sum_i wL_i - sK$ is the regulated rate-of-return constraint):

$$\max Z = \sum_{i} R_i - \sum_{i} \frac{wQ_i}{f(k)} - rK + \sum_{i} v_i \left(K - \frac{kQ_i}{f(k)} \right) - \mu \left[\sum_{i} R_i - \sum_{i} \frac{wQ_i}{f(k)} - sK \right]. \tag{10}$$

This utility chooses Q_i, K, and k:

$$\frac{\delta Z}{\delta Q_i} = \left(MR_i - \frac{w}{f(k)} \right) (1 - \mu) - v_i \frac{k}{f(k)} = 0, \tag{11}$$

$$\frac{\delta Z}{\delta K} = (\mu s - r) + \sum_t v_t \leq 0, \tag{12}$$

$$\frac{\delta Z}{\delta k} = \frac{wQ_t}{f^2} f_k(1 - \mu) - \frac{v_t Q_t}{f^2}(f - kf_k) = 0, \tag{13}$$

$$\frac{\delta Z}{\delta v_t} = K - \frac{kQ_t}{f(k)} \geq 0, v_t \geq 0, \tag{14}$$

$$\frac{\delta Z}{\delta \mu} = \sum_t R_t - \sum_t \frac{wQ_t}{f(k)} - sK = 0. \tag{15}$$

From (11) and (12) the peak and off-peak prices of this regulated utility can be estimated[6] (an asterisk denotes the optimal value of the variable for the regulated utility):

off-peak:

$$MR_j^* = w \frac{L_j^*}{Q_j^*}; \tag{11a}$$

peak:

$$MR_p^* = w \frac{L_p^*}{Q_p^*} + \frac{v_p^*}{(1 - \mu)} \frac{K^*}{Q_p^*}. \tag{11b}$$

Summing over all peak periods (and substituting for $\sum_t v_t$ from [12]) yields

$$\begin{aligned} \sum MR_p^* &= \sum_p \left(w \frac{L_p^*}{Q_p^*}\right) + \frac{r - \mu s}{1 - \mu} \frac{K^*}{\sum_p Q_p^*} \\ &= w \sum_p \left(\frac{L_p^*}{Q_p^*}\right) + r \frac{K^*}{\sum_p Q_p^*} - s^1 \frac{K^*}{\sum_p Q_p^*}, \end{aligned} \tag{11c}$$

where $s^1 = [\mu(s - r)]/(1 - \mu) > 0$. Summing over all off-peak periods ($v_j^* = 0$) yields

$$\sum MR_j^* = w \sum_j \left(\frac{L_j^*}{Q_j^*}\right). \tag{11d}$$

Bailey attempts to show that off-peak users are charged the same price under regulation as they would be if the firm was unregulated. Only peak users benefit from regulation, since to them alone is the price reduced. These conclusions and Bailey's proof of efficiency in production depend on the assumption that $L_j/Q_j = L_j^*/Q_j^*$, which is true only under fixed proportions.[7] Before establishing that off-peak users also benefit from regulation, we must determine the impact of regulation on the choice of the technique of production.

---

[6] It is assumed that $0 < \mu < 1$. Regulation is binding; therefore, $\mu \neq 0$. The type of argument used by Bailey would show $\mu < 1$. More generally, an analysis of the second-order condition would show $\mu < 1$ if $R_{LL} < 0$ (the second derivative of revenue with respect to labor is negative).

The capital-labor ratio k is determined before capital is fixed in place but after relative factor prices and *the presence of regulation* are considered.

By rearranging the terms in (11c), we can determine the equation indicating the capital-labor choice for the regulated utility:

$$\frac{\sum Q_p^* f_L}{\sum Q_i^* f_k} = \frac{(1 - \mu)w}{r - \mu s} = \frac{w}{r} + \frac{w\mu}{r}\left(\frac{s - r}{r - \mu s}\right), \qquad (16)$$

where $s - r$ is positive as given and, from (12), $r - \mu s$, the contribution of capital, is positive at the peak ($v_i > 0$) and zero off peak ($v_i = 0$). The second term in (16) is therefore positive. Compare (16), the equation showing the capital-labor choice of the regulated utility, with (9), the same equation for the unregulated monopoly. For a binding regulatory constraint ($\mu \neq 0$) and $\mu \neq 1$, the regulated utility sets the weighted ratio of marginal products of labor to capital to something *more* than the ratio of factor prices. Where f_L and f_k are both positive but f_{kk} and f_{LL} both negative, this can only be achieved by increasing the capital-labor ratio. Therefore $k^* > k$. The regulated utility operating with a fixed technique of production in the short-run ex post case but which can vary capital and labor freely, constrained only by the production function and regulation, in the long run ex ante case chooses an inefficiently large capital-labor ratio for all levels of output, peak and off peak. Equation (16) is analogous to the "standard" description of inefficiency modified by two factors to incorporate the specific ex ante putty, ex post clay model (Baumol and Klevorick 1970, p. 166). First, the term $r - \mu s$ represents the contribution of capital at the peak only. Second, adjustment is made to the ratio of marginal products reflecting the percentage of time which is peak.

Now, compare the off-peak price (and therefore output) choices for the unregulated and regulated firms:

unregulated:

$$MR_j(Q_j) = \frac{wL_j}{Q_j} = \frac{w}{f(k)} \qquad \left[\text{from } \frac{Q_j}{L_j} = f(k)\right];$$

regulated:

$$MR_j(Q_j^*) = \frac{wL_j^*}{Q_j^*} = \frac{w}{f(k^*)}.$$

---

[7] In the first part of the paper, it is shown that Bailey implicitly used a fixed-proportions production function where $q_i/L_i = 1/a$ and a is a constant.

Under the assumptions of a monotonic declining marginal revenue curve, positive marginal products of capital and labor, and the above proof that $k^* > k$,

$$MR_j(Q_j^*) < MR_j(Q_j) \qquad \text{since } f(k^*) > f(k). \tag{17}$$

The regulated utility chooses a higher average capital-labor ratio in all periods as compared with the unregulated monopoly. As a result, assuming reasonable production characteristics, off-peak output increases when regulation is imposed. Given downward-sloping demand and marginal revenue curves, the price to off-peak users falls. Regulation benefits all users, peak and off-peak.

Summary

In a model where the firm chooses ex ante the capital-labor ratio (technique of production) it will use over a number of periods, the choices of capital and labor are biased by regulation on capital's rate of return. In general, where some capital-labor substitution is also available ex post, additional incentives for capital bias would exist. The conclusion postulated by Bailey that peak-load pricing models do not indicate such bias is untrue even in the explicit model used by her if that model is considered to be only one segment of the utility's decision space. Operating with a fixed-proportions production function in the short run will involve inefficiency if there is long-run substitutability between capital and other factors. Unless it can be shown that the elasticity of substitution is zero between capital and other factors of production for the technological choices available to a utility in the long run, simply adding more periods to the single-period *A-J* model cannot change its most important conclusions.

This capital bias has important effects on the pricing policies of the regulated firm, peak and off peak. The existence of a capital-labor ratio greater than that dictated by market prices of factors leads to the firm's expansion of output in *all* periods. Both peak and off-peak users benefit from rate-of-return regulation.

References

Atkinson, A. B., and Waverman, L. "Resource Allocation and the Regulated Firm: Comment." *Bell J. Econ. and Management Sci.* 4, no. 1 (Spring 1973): 283–87.

Bailey, E. E. "Peak-Load Pricing under Regulatory Constraint." *J.P.E.* 80, no. 4 (July/August 1972): 662–79.

Bailey, E. E., and Malone, J. C. "Resource Allocation and the Regulated Firm." *Bell J. Econ. and Management Sci.* 1, no. 1 (Spring 1970): 129–42.

Baumol, W. J., and Klevorick, A. K. "Input Choices and Rate of Return Regulation: An Overview of the Discussion." *Bell J. Econ. and Management Sci.* 1, no. 2 (Autumn 1970): 162–90.

Littlechild, S. "Marginal Cost Pricing with Joint Costs." *Econ. J.* 80, no. 318 (June 1970): 323–35.

McNicol, D. L. "The Problem of Peak Load Pricing Subject to Rate of Return Constraint," mimeographed (Philadelphia: Univ. Pennsylvania, November 26, 1973).

Williamson, O. E. "Peak Load Pricing and Optimal Capacity under Indivisibility Constraints." *A.E.R.* 56, no. 4 (September 1966): 810–27.

"An Evaluation of Consumer Protection Legislation: The 1962 Drug Amendments": A Comment

Thomas McGuire, Richard Nelson, and Thomas Spavins
Yale University

In a recent article in this *Journal*, Peltzman (1973*a*) develops and implements a procedure to measure the losses to consumers that result from their lack of full knowledge regarding the quality of a new product. He then uses this procedure to assess the benefits and costs of the 1962 amendments to the Food, Drug, and Cosmetics Act. The article has received a considerable amount of attention, and the conclusions and methodology have been touted in the popular press (Friedman 1973). While we admire Peltzman's courage and ingenuity, we believe that his methodology is fundamentally flawed in several respects. We are concerned, therefore, that the conclusions are being given more credence than they are worth, and, more important, that a basically faulty methodology will be employed in future studies of this genre.

The problems with the Peltzman article are at two levels. Many economists and analysts of the medical care system would have doubts about the straightforward application of consumer demand theory to the problem in question. However, the critique we present here of the Peltzman paper is not at this fundamental level, but rather of the particular procedures he follows in carrying out an analysis, based on traditional consumer demand theory, of the cost of consumer misinformation regarding product quality.[1] His procedures are as follows.

Since much of the discussion preceding the drug amendments was concerned with misinformation leading consumer-physicians to purchase new drugs when old drugs were as efficacious and cheaper, Peltzman focuses on this case. He assumes that, initially, as a result of the consumer's overevaluation of the new product, his demand curve is *ADM* in figure 1.[2] The demand curve *GHEN* refers to the situation where the new drug is

---

[1] We shall not discuss here the problems introduced to the analysis in assuming that doctors make the choices among various drugs at various prices that their patients would have made had they known what the doctors know. The demand curve in fig. 1 will simply be treated like any consumer demand curve.

[2] See n. 1 above.

[*Journal of Political Economy*, 1975, vol. 83, no. 3]
© 1975 by The University of Chicago. All rights reserved.

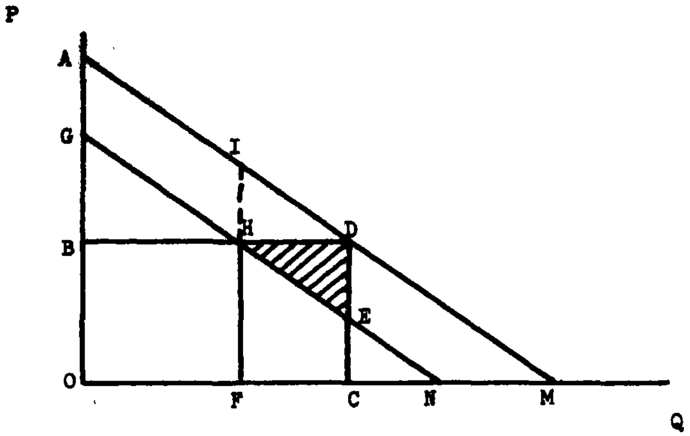

FIG. 1

correctly evaluated. Given a fixed price, *OB*, Peltzman argues that triangle *HDE* represents net loss per period to the consumer due to his ignorance, and that it is a "welfare" cost analogous to triangle losses associated with excise taxes, etc. Peltzman assumes that, as the consumer acquires experience with the product, he learns to evaluate it correctly and his actual demand curve shifts over time from *ADM* to *GHEN*. The total loss to the consumer due to his ignorance is the sum over time of the (diminishing) triangle losses properly discounted. For expositional purposes, Peltzman considers the case in which there are only two periods, in the first of which the consumer does not know and in the second of which he knows fully actual product quality. Then the total loss is triangle *HDE*. In our discussion we will work with this simpler specification of the problem.

In applying the model to assess the benefits of an event (legislation) designed to increase new product information to consumers, Peltzman proposes that the effects should be measured in terms of differences between the triangle losses before and after the legislation. Peltzman argues that, in the example of figure 1, improved information and its value should show up in terms of a smaller falloff of demand by a particular consumer for a new product in the period after it is introduced (and hence smaller triangle losses), since he will initially make less of an overestimate regarding product quality.

Peltzman moves from this exposition of the losses experienced by a single hypothetical consumer under two different information regimes, to measurement of total losses suffered by all consumers, using market demand curves. He estimates the first initial or "before-learning" market demand curve for a new drug, and then, at some later time after consumers have gained experience with the drug, the objective or "after-learning" market demand curve. These two demand curves are analogous to *ADM* and *GHEN*, respectively, for the single consumer. This is done for new drugs introduced in the pre- and postamendment periods. The effects (if any) of better consumer information should show up, Peltzman argues, in market demand curves in the postamendment period which fall off less sharply after introduction of the product, and hence in smaller triangle losses estimated from those market demand curves. This is how Peltzman derives his estimates of the "benefits "of the legislation.

We have three objections to Peltzman's methodology for measuring the benefits of the drug amendments. The first has to do with Peltzman's assumption that all consumers make the same kind of error, in particular that they overestimate the value of a new drug.[3] While this type of error may predominate where the source of consumer misinformation is inflated manufacturer claims about product effectiveness, this is surely not the only type of error that will occur. In the absence of adequate physician-patient knowledge regarding drug efficacy, some physicians treating some ailments will not prescribe new drugs when they should, perhaps because of doubts about the claims, etc. This second class of consumers also will experience a triangle loss, but *IHD* rather *HDE*. These two kinds of losses could be added if they could be separately observed. But they cannot be inferred from shifts in market demand curves. While, as the result of learning, the first class of consumers will shift their demand curve to the left, the second class of consumers will shift their demand curve to the right. As a result of learning, the aggregate demand curve may shift to the left or shift to the right, or it may not move at all. There is no way to get the sum of triangle losses by observation of the aggregate demand curve and its shifts over time. And, if consumer mistakes are of both sorts, there is no logical connection between the advent of better information regarding a new product and subsequent changes in the way the aggregate demand curve for a new product shifts over time. Consider two different regimes, one in which there was considerable early ignorance on the part of consumers, the other in which all consumers were well informed from the beginning. In these two radically different cases, one could observe the same time path of demand for a new product in the years after it had been introduced.

Our second objection has to do with the presumed analogy between the welfare effects of an excise tax and consumer misappreciation of product quality. The analogy is workable in the care of a single individual but breaks down completely in the movement so market demand functions. Even if all consumers were to make the mistake of overevaluating drug effectiveness, there is no reason to believe that they would all make errors of the same magnitude. And if they do not, the sum of the individual triangle losses will not equal (in general will be greater than) triangle losses estimated from data from aggregated market demand curves.[4]

---

[3] Peltzman recognizes that there may be two different kinds of error but then ignores the problem (1973*a*, p. 1062).

In the usual case where we estimate loss of consumer surplus using data from market demand curves, it is assumed that all consumers of the product face the same distortion, say an excise tax like IH. Let $IH = DE = \Delta P_i$, and $HD = \Delta Q_i$. The sum of losses to consumers is

$$W = -\tfrac{1}{2} \sum_i \Delta P_i \, \Delta Q_i.$$

If $\Delta P_i = \Delta P$ for all i, the sum can be expressed as

$$W = -\tfrac{1}{2} \Delta P \sum_i \Delta Q_i = -\tfrac{1}{2} \Delta P \, \Delta Q.$$

The ΔP and ΔQ refer to observed aggregative shifts and can be associated with points along a market demand curve. But this calculation does not go through if each consumer faces a different price distortion or if (as in the case of Peltzman) the amount of overvaluation may differ from consumer to consumer. The key operation of moving the ΔP_i through the summation sign then is illegitimate.

Consider a two-person market and two different circumstances. In the first case both consumers make the same overestimate, IH, of marginal value (schedule) when the product is first introduced, while in the second circumstance one of the consumers judges exactly correctly but the other makes twice the error ($2IH$) of the first circumstance. Assume that the two individuals have identical slopes of their demand curves. Then, since the loss to a consumer is not linear but a quadratic function of his overvaluation (price distortion), the total losses in the second case are greater than in the first. But if one looks at market data when the product is new, the total quantity purchased (at any price) will be the same under the two circumstances. If the "market demand" curve is observed again after consumers have learned the facts, the same "backward shift" will be observed in the two different situations. Thus the losses calculated from market data will be the same in both cases. But these will equal the sum of consumer losses only under the first assumption—that all consumers make the same overevaluation (face the same price distortion).[5]

---

[4] This problem is familiar from discussions of the welfare cost due to monopoly (see, e.g., Scherer 1970, p. 403).

[5] In general, the estimate of loss of consumer surplus using market demand data will be smaller than actual consumer loss, except in the case where all of the ΔP_i are the same. Let $\Delta Q_i = S_i \Delta P_i$, where s_i is the inverse of the slope of the individual's demand curve. Consider a distortion characterized by an (unobserved) vector of possible different "taxes" or bad estimates of quality, $\Delta P_1 \ldots \Delta P_i \ldots \Delta P_n$, which yield an (observed) total change in quantity $\Delta Q = \sum_i \Delta Q_i$ from the optimal situation. Then

$$W = -\tfrac{1}{2} \sum \Delta P_i \, \Delta Q_i = -\tfrac{1}{2} \sum \Delta P_i^2 S_i. \tag{1}$$

There is a third problem which, although it probably does not pertain to all cases in which one is estimating the costs of product misinformation, certainly does pertain to the particular case analyzed by Peltzman. Recognize that when a doctor decides to prescribe one kind of a drug this generally is an implicit decision not to prescribe another. Assume that the treatment with the drug does nothing for the patient and treatment with another drug would have done something. We will exclude from consideration, as does Peltzman, new drugs that actually are harmful.[6] Cost to the patient of the incorrect decision includes the Peltzman triangle, but there is an additional element of cost: the value to the patient of treatment with the old drug which was foreclosed by the wrong decision to use the new one. In principle, one could have obtained the total loss of the incorrect decision by obtaining from each past user of the drug a statement regarding what he would have paid to procure and use the drug (to the exclusion of other drugs) had he known better. But that amount well might have been negative. And one does not obtain that kind of data from observing actual market demand curves.

In putting forth these three objections, our central point is that the researcher may not rely on aggregate demand relations for the information necessary to assess costs of consumer misinformation. These issues suggest caution in interpretation of Peltzman's results; in particular, all three

---

This loss will depend on the vector of (price) distortions that generated the observed quantity distortion. To find the set of ΔP_i which imply *minimal* W for a given ΔQ, differentiate the following Lagrangian with respect to each ΔP_i:

$$W(\lambda) = -\tfrac{1}{2}\sum \Delta P_i^2 S_i + \lambda\left(\sum \Delta P_i S_i - \Delta Q\right); \tag{2}$$

$$\frac{\partial W(\lambda)}{\partial \Delta P_i} = -S_i\,\Delta P_i + \lambda S_i. \tag{3}$$

W clearly has an extreme value where $\Delta P_i = \lambda$ for all i. Consideration of the overall shape of the function reveals that this point is a unique minimum. If the ΔP_i are not all the same, W will be greater than this. If the loss of consumer surplus is calculated from "market" data, one can visualize the calculations as follows. It will be observed that $\Delta Q = \sum \Delta Q_i$. The slope of the estimated market demand curve will be $1/\sum S_i$. The loss of consumer surplus then will be calculated as

$$\hat{W} = -\tfrac{1}{2}\Delta Q\,\Delta\hat{P} = -\tfrac{1}{2}\Delta Q^2/\sum S_i. \tag{4}$$

The "hat" over W and ΔP refers to the implicit presumption in the calculation (which may be counter-factual) that all consumers face the same distortion. One can rewrite "actual" consumer surplus as

$$W = -\tfrac{1}{2}\sum \Delta Q_i^2/S_i. \tag{1a}$$

Thus

$$W = \hat{W}\left[\left(\sum \frac{\Delta Q_i^2}{S_i}\right)\Big/\left(\frac{\Delta Q^2}{\sum S_i}\right)\right]. \tag{5}$$

One can calculate that $W = \hat{W}$ if $\Delta Q_i/S_i = \Delta P_i = \lambda$ for all i. Then the numerator and the denominator in the brackets will be equal. And from (3) it is apparent that $W > \hat{W}$ if this condition is not met, since the numerator is a unique minimum under this condition.

[6] Peltzman discusses the hazard problem in another paper (1973*b*).

objections suggest Peltzman had underestimated gains from increase in consumer information.

Other difficulties, in addition to these, flaw Peltzman's attempt to measure the costs of the new legislation. In the debates about the amendments, it was recognized that there would likely be some costs associated with the tighter screening of new drugs that would increase the expense of introducing new drugs and occasionally screen out a new drug that should be introduced. Peltzman's method of measuring these costs is not satisfactory.[7]

His analysis of the costs of restricting entry of new drugs falls into two partially connected parts. The first is his analysis attempting to demonstrate that the new legislation curtailed drastically the rate of introduction of new drugs. While one may question whether the entire post-1962 decline in new drug introduction is attributable to the amendments (and the variables in his model), we certainly believe that the legislation did cut back the number; this was one of its intents. The hope of the proponents was that the drugs that would not be available would be largely ones that were not effective or which were at best marginal improvements, and that the bulk of the effective ones would be available. Peltzman argues, we think plausibly, that if this hope had materialized, one might have expected a falling off in the share of total drug sales accounted for by new drugs after the amendments, but by less than the percentage reduction in the rate of introduction of new entities. The percentage decline in sales is indeed somewhat less than the percentage decline in new drug entries, but not by very much. The question is, To what does one attribute the large postamendment decline in sales of new drugs relative to sales of total drugs?

Peltzman argues that as the price of new drugs did not change much, their decline in relative importance must be attributed to a backward shift in the demand curve for new drugs relative to established drugs. He then argues that his analysis has shown that the backward shift cannot be the consequence of better-informed consumers less prone to exaggerate the merits of new drug entities. Next he makes a curious proposition: that the backward shift in the demand curve was the consequence of restrictions, imposed by the amendments, on the consumers ability to learn about new drugs, "reduced information content associated with new drugs" (1973*a*, p. 1071). He treats the backward shift in the demand curve for new drug entities as an artificial restriction on demand, and proceeds to estimate the loss of consumer surplus. Peltzman seems to be trying to have the argument both ways at once. If, as Peltzman argues, information regarding new drug entities was artificially restricted in the postamendment period, one could have expected that demand for the new drug would grow more (or decline less) as a result of experience than would have been the case in the preamendment period. But, of course, Peltzman earlier argued that changes in demand for a new drug over time were not significantly different in the pre- and postamendment periods.

---

[7] Peltzman seeks to measure these costs only insofar as they affect consumers. His analysis of the impact of amendments on pharmaceutical producers is developed separately (1973*b*).

Peltzman's awkward "restriction of information" argument may well have been forced upon him by the expedient he took in his econometric work on the demand for new drugs, in which he distinguished between new and established drugs but assumed that all new drugs were homogeneous. The most plausible case for Peltzman to argue is that the legislation resulted in a screening out of some new drugs which would have been efficacious and for which other drugs, new or old, were not perfect substitutes. This would, indeed, show up in a backward shift in the demand curve for the group of new drugs taken as a whole. But Peltzman's econometric analysis does not permit him to pose the problem in this manner. And since his argument that demand could not have fallen off because consumers were being duped less is unpersuasive, the appropriate attribution of the demand falloff remains uncertain. Further, the costs to consumers of restriction on entry of certain new drugs that are not perfect substitutes for other new drugs cannot be calculated by observing shifts in the total "demand curve" for new drugs.

This note is not intended as a general critique of the Peltzman article. As suggested earlier, one can criticize the article at a deeper level than we have here. On the other hand, there is much in the article which is well conceived and which probably points in the right direction. Nor is this note intended to argue against Peltzman's conclusions regarding the legislation he is studying. He may be right: perhaps the bill should be repealed. But Peltzman's analysis provides no real evidence on which one can rest a judgment of this sort. And the methodological problems with his analysis suggest that the answer cannot be found simply by looking at what has happened to market demand curves.

References

Friedman, Milton. "Frustrating Drug Advancement." *Newsweek* (January 8, 1973), p. 49.

Peltzman, Sam. "An Evaluation of Consumer Protection Legislation: The 1962 Drug Amendments." *J.P.E.* 81, no. 5 (September/October 1973): 1049–91. (*a*)

———. "The Benefits and Costs of New Drug Regulations." In *Regulating New Drugs*, edited by Richard Landau. Chicago: Univ. Chicago Center Policy Study, 1973. (*b*)

Scherer, Frederick M. *Industrial Market Structure and Economic Performance*. Chicago: Rand-McNally, 1970.

AUTHORS' NOTE. It is unfortunate that Peltzman read our note as an essay in advocacy of present FDA mandates and policies. Not so at all. Rather, our purpose was to point out that Peltzman's methodology was fundamentally flawed. In his reply Peltzman presents nothing that detracts from the force of our criticisms. In fact, he attempts to blunt our accusation of bias by conceding that his after-learning demand curve may not measure the true consumer valuation of new drugs. But then, his estimates of benefits and costs are meaningless.

"An Evaluation of Consumer Protection Legislation: The 1962 Drug Amendments": A Reply

Sam Peltzman

Graduate School of Business, University of Chicago

I am grateful to McGuire, Nelson, and Spavins (MNS) for unwittingly strengthening the case I made in my article that the costs of the 1962 amendments to the Food, Drug, and Cosmetic Act exceed the benefits. The burden of their criticism is that I ignored some potential benefits of the amendments. No mortal would pretend otherwise, but MNS's specific points also imply that I have ignored some costs as well. I will argue that, on balance, the MNS critique implies that my admittedly crude estimate of the net costs of the amendments is biased downward.

I will confine my elaboration to the three main points raised by MNS:

1. MNS argue that dissipation of misinformation provides a gain even if the demand for a good remains unchanged. This would occur if the misinformation consisted of equal amounts of over- and underevaluation of the good. I have no quarrel with this point in principle, but its application to the 1962 amendments is highly dubious. In the unregulated, preamendments drug market, producers already had a private incentive to reduce underevaluation of drugs. In addition, they were alleged to have a private incentive to produce overevaluation, through extravagant advertising claims. The amendments were therefore designed specifically to cure the overevaluation problem, and, in the process, they work to exacerbate the underevaluation problem. Some advertising that might reduce underevaluation will now be rationed out by the cost of complying with the amendments' proof-of-efficacy requirement. This is precisely the kind of force which will generate, not eliminate, the loss *IHD* described by MNS. An important example, among many others cited by Wardell (1974), concerns the drug propranolol, which is used widely abroad for management of angina pectoris and hypertension. Its use for these maladies is also advocated by some American texts and the American Medical Association. However, the FDA has not yet approved this particular application, so advertising of it remains prohibited. In consequence, some doctors who would have prescribed the drug in the absence

[*Journal of Political Economy*, 1975, vol. 83, no. 3]
© 1975 by The University of Chicago. All rights reserved.

of the amendments are not doing so, and a loss of the type that MNS claim is reduced by the amendments is in fact generated by them.

MNS then face a difficult task. They must argue that a regulatory process which taxes the production of information that might reduce underevaluation of new drugs somehow ameliorates underevaluation. How do they do this? The only hint we get is an implicit assertion that the amendments work to strengthen doctor confidence in advertised claims "perhaps because of doubts about the [unregulated] claims, etc." Even if we ignore the role of such doubt in limiting an unregulated seller's production of exaggerated claims, MNS leave too much unsaid. The very fact that advertising of some useful applications of a drug is rationed out by costs of complying with the amendments will sow its own doubts—about the validity of unadvertised uses. Doctors who hear of these uses will be made uncertain about them by their lack of official sanction (and the potential this raises for malpractice suits). The more general point is that the amendments produce some information which can reduce uncertainty, but also inhibit production of other similar information I cannot see how MNS can deduce a net reduction of uncertainty a priori.

Since the issue must be resolved empirically, I will venture a modest extension of the evidence in my article about the effect of the amendments on consumer confidence. (I argue there that increased confidence should reduce the cross-drug variability of sales growth, but there is no evidence that this has occurred since the amendments [Peltzman 1973*a*, p. 1078, n. 30].)

One measure of physicians' confidence in their evaluation of a new drug is willingness to prescribe refills of an initial prescription. A physician can dispel some uncertainty about the value of a drug by prescribing it for a likely malady in order to observe the effects on the patient. If this initial trial is unsuccessful, the initial prescription will not be refilled. Where the drug is successful, the initial trial will be followed by refills where these would be indicated in treatment of the malady. In a market characterized by much uncertainty, or many uncertain physicians, there will be an atypically high proportion of initial prescriptions—coming from doctors who are groping for both the appropriate and inappropriate uses for a new drug. If the amendments work to increase physician certainty through the added information they generate, the demand for this physician experimentation should be lower (and the probability of a successful initial trial, greater). Doctors will be equipped to zero in on the valid uses quickly, so that the new drug will quickly attain its equilibrium refill proportion.

Unfortunately, the data require us to look at 3-year-old rather than brand new drugs, but these data do not support the notion that the amendments have increased physician confidence in their drug evaluations. I calculated the ratio of refills to initial prescriptions (r) in the third year after introduction for drugs introduced in 1961–62 and for those introduced in 1965–67.[1] Now, r will vary among drugs for reasons unrelated to the point at issue. For example, r will be higher for longer-term maladies, so it will be lower for antibiotics than for tranquilizers. To adjust for this, I divided r by the same ratio ($\bar{r}$) for the aggregate of all other drugs in the same "therapeutic category" (roughly, a group of drugs treating similar conditions; the data and categorization are from R. A. Gosselin & Co. [1971].) This normalized $r(R)$ will be less than one if doctors are more uncertain about new than old drugs. For the 1961–62 sample (23 drugs), the mean (standard error) of R is .97 (.07); for the postamendment sample (26 drugs), R is .92 (.09). If anything, this "confidence index" is lower for postamendment drugs, though the difference is insignificant. Alternatively, one may regress the log of r on the log of $\bar{r}$ and a dummy variable (A) equal to one for the postamendment drugs. The result is (t-ratios are below coefficients)

$$\begin{array}{rcccc} \ln r = & -.04 & + .99 \ln \bar{r} & - & .18A, \\ & (-.05) & (6.94) & & (-1.00) \end{array}$$

which implies a roughly 16 percent lower value of r for the postamendment drugs, but again the difference is insignificant.

These data are hardly elegant or conclusive—for example, we would really like to know the willingness of the doctor to initiate and refill prescriptions for other patients after a few early patient trials—but they do illustrate the risk in merely asserting a potential benefit of the amendments while ignoring offsetting costs.

2. MNS point out that the implicit analogy I drew between the welfare effects of a tax and those of misestimation of demand is imprecise in the aggregate, because individuals will make different misestimations. This is correct in principle, but the resulting bias hardly goes one way and is likely to be conservative in this case. The aggregation procedure I used will be inaccurate whenever the gap between initial and true evaluations of demand differs among consumers. MNS consider a particular case where the aggregate initial demand (*ADM* in their fig. 1) is an average of one individual's correct evaluation of the drug (*GHEN*) and another's wildly optimistic evaluation (a demand laterally in excess of *ADM* by *FC*). In this case 2*HDE* underestimates the potential gain of the amendments. But the amendments also restrict production of information, so we must consider the complementary possibility: this restriction differs among individuals, so both initial evaluations in the MNS example would

---

[1] Specifically, the drugs are major new chemical entities, where "major" means that they took at least 1 percent of prescriptions in a therapeutic category that accounted for over 1 million prescriptions annually. The sample is discussed more fully in Peltzman (1973*a*).

have been accurate without the amendments. In this case, the amendments' restriction of information is nil for one individual and larger than is implied by the shift of the aggregate demand for the other. In this case, my estimate of the cost of the amendments ($2AGHD$) is too low by a factor involving the square of FC: in the limit where the drug in question is kept off the market forever, the true cost of the amendments is twice my estimate. The comparison of the behavior of the aggregate demand curves over time in my article implies strongly that this downward bias in the estimate of the cost of the amendments is many orders of magnitude more important empirically than the downward bias in the benefits of the amendments that MNS chose to single out. Of course, we do not know the absolute amounts involved because we do not know the dispersion of individual drug demands. This is why I had to implicitly assume a "representative consumer." But I am grateful to MNS for lending further support to my contention that my estimate of the net costs of the amendments is conservative.

3. MNS argue that my technique assumes a zero or positive value for both old and new drugs and that, in consequence, I exclude losses from new drugs which have negative value. I have treated one aspect of this problem elsewhere (Peltzman 1973*b*), but not that cited by MNS—the possibility that the use of a new drug forecloses successful treatment with old drugs (and the new drug demand does not discount this possibility). Again, however, MNS have failed to consider the other side of their coin: some old drugs will have a negative value compared with new drugs which the amendments have foreclosed. This can occur ex ante (the old drug is too risky, unpalatable, etc., so it is not prescribed) or ex post (the old drug turns out to harm more than it helps). If the negative value of old drugs is also imperfectly reflected in the demand for new drugs, my procedure understates the benefits of new drugs. I cannot claim to know whether the new drug benefits I have thereby excluded are more or less than the costs I have omitted. However, Wardell (1973, 1974) implies strongly that the cases in which old drugs have essentially negative value compared with new drugs (in terms of added hospitalization costs, side effects, etc.) far outnumber those in which new drugs have negative value.[2] This may not be surprising, since drug sellers will both bear at least some of the negative value of new drugs (their reputations will be damaged) and collect some of the negative value of old drugs from sales of new replacements.

In short, the main shortcoming of the MNS critique is its consistent failure to contemplate all the relevant implications of their analysis. None of the omissions they point to has the one-sided implications they allege, and a symmetrical treatment of them serves mainly to reinforce my conclusion that the costs to consumers of the 1962 drug amendments have far exceeded the benefits.

---

[2] To cite one example, Wardell (1974) estimates that delay in FDA permission for marketing a hypnotic that was safer than available alternatives cost 1,200 lives in the United States.

References

Peltzman, S. "An Evaluation of Consumer Protection Legislation: The 1962 Drug Amendments." *J.P.E.* 81, no. 5 (September/October 1973): 1049–91. (*a*)

———. "The Benefits and Costs of New Drug Regulations." In *Regulating New Drugs*, edited by Richard Landau. Chicago: Univ. Chicago Center Policy Study, 1973. (*b*)

R. A. Gosselin and Co. *National Prescription Audit, Therapeutic Category Report, Ten Year Trend, 1961–70.* Dedham, Mass.: Gosselin, 1971.

Wardell, W. "British Usage and American Awareness of Some New Therapeutic Drugs." *Clinical Pharmacology and Therapeutics* 14, no. 6 (November/December 1973): 1022–34.

———. "Therapeutic Implications of the Drug Lag." *Clinical Pharmacology and Therapeutics* 15, no. 1 (January 1974): 73–96.

Book Reviews

The Economic Theory of Representative Governments. By ALBERT BRETON. Chicago: Aldine-Atherton, Inc., 1974. Pp. xvi+228.

This book consists of an application of the theory of supply and demand to the output of public goods by representative governments. Since this is a quite new way to discuss governmental activity, the book contains a remarkable number of deep and original insights about heretofore unrecognized parameters of political behavior. At the same time, the theory yields few predictions of outcomes and for that reason may very well lose out in competition with approaches that are more fruitful though perhaps less theoretically rigorous.

To outline the main elements of the argument, on the demand side, it is assumed that citizens have preferences for policies (not for politicians) and want to maximize their (citizens') welfare relative to the net benefits (i.e., net of tax costs) from public goods. Owing to the lumpiness of these goods, however, nearly all citizens are coerced into accepting either more or less than they would ideally like to have of a particular public policy. Such coercion is held to be the mainspring of demand: the greater the degree of coercion suffered by a citizen, the greater is his participation (by, e.g., lobbying, voting, etc.) in political life.

On the supply side, producing policies or outputs are two kinds of actors: governments (consisting of politicians or representatives) and bureaucrats (who are only partially responsive to representatives). Their motives are assumed to be, for governing politicians, to maximize their chance of being reelected (in spite of criticism which is the motive and function of the opposition), and, for bureaucrats, to maximize the chance of increasing the budgets of bureaus. In the actual supply of public goods, the governing part attempts to reduce the degree of coercion in public policies in order to gain the support of citizens for the party's reelection. In so doing, the government must work against the criticism of opposition and the greed of bureaus.

Actual resource allocation flows from the forces of demand and supply thus: Citizens with given tastes are coerced (by the lumpiness of policies) and respond by participating in efforts to influence the representatives. They respond in turn by attempting to adjust policies to minimize coercion (i.e., by revision of policies and negotiation with bureaucrats). Three cases are distinguished according to the cost of participation and the ratio of the number of people participating to the number needed to win. If the cost is low and this ratio is therefore less than one, politicians have some freedom of action in producing policies. If the cost is high and the ratio greater than or equal to one, politicians are constrained to satisfy citizens' preferences as exactly as they can.

As can be seen from this outline, the author's argument contains a number of highly original and persuasive notions, for example, that participation is a negative reaction to coercion (instead of, as is usually argued, a direct striving for benefits) and that representatives have greater freedom of action when the number of participants is especially large. Another attractive distinction is found in the demonstration that goals (or policy outputs) and instruments of achieving them are somewhat interchangeable, as when maximizing the number of policemen is substituted for the policy of police protection. There are some errors in these subsidiary arguments, as when the author relies on logrolling to produce a rational party policy even though it has been repeatedly demonstrated that logrolling involves voting paradoxes. Still, in spite of a few such perversities, the book is full of enlightening and intuitively attractive notions.

Nevertheless, it must be judged on the force and attractiveness of its main argument. And here there is a weakness. In spite of the excellent description of the forces of supply and demand, this description does not lead to a discussion of the actual form of political equilibria, that is, to a discussion of the actual allocation of policies. Observing the uses of the theory of supply and demand to discuss economic equilibria and even prices, one might expect to find a discussion of policies adopted or a least of categories of policies. Instead the analysis falters, ending with an inconclusive discussion of, for example, relative bargaining strengths of politicians and bureaucrats.

In sharp contrast with this failure, an alternative theory now (1974) stands on the verge of some success in describing equilibria in public policy. This is the theory of duopoly, also adapted (by Anthony Downs) from economics. Although Breton dismisses it rather cavalierly, it has in the past decade been rendered both more powerful and more realistic by extending it to two or more dimensions (Davis and Hinich) and by accounting for kinds of institutional variations (Ordeshook and McKelvey). It is now at the point that one can subject it to empirical verification using data from survey research (Hinich and Rabinowitz) and even perhaps to testing by prediction. This spatial model of competition between parties to supply policies attractive to citizens leads to a description of an equilibrium of policies in terms of the demands of (usually) the median voter. This is a much more concrete description than appears possible from Breton's analysis, and it therefore seems likely that in the long run the spatial duopoly model will turn out to be more attractive than Breton's.

William H. Riker

University of Rochester

Cost-Benefit Analysis: Theory and Practice. By Ajit K. Dasgupta and D. W. Pearce. New York: Barnes & Noble, 1972. Pp. 270. $13.75.
Cost Benefit and Cost Effectiveness: Studies and Analysis. Edited by J. N. Wolfe. London: George Allen & Unwin, 1975. Pp. 236. £4.50.

Dasgupta and Pearce discuss principles, whereas the Wolfe volume is a collection. I discuss the latter in light of the central issue raised by the former.

Dasgupta and Pearce begin with the gap they see between welfare theory and its application in cost-benefit analysis (CBA). They discuss the relationships among preferences, ordinal and cardinal utility, and "willingness to pay," problems in making social utility operational, the use of demand and supply curves in measurement, Pareto optimality and compensation tests, and the relevance of Arrow's "possibility theorem," and conclude that CBA is essentially inconsistent because its aggregate measures of preference violate "reasonableness." This result is troublesome for their subsequent investment criterion of ranking projects by the maximand of benefits minus costs. Further, it leads them into ambiguity about whether CBA should be a guide to, or a substitute for, political decisions. Their text is written in terms of the former, but part 1 and their implicit theme assume that CBA *ought*, in some sense, to measure social utility, though they cannot deal with income distribution aspects.

Part 2 of *Cost-Benefit Analysis* deals with accounting prices, part 3 deals with decision formulae, and part 4 presents two case studies. The treatment of accounting prices is wholly unsatisfactory. Nine cases of divergence between market and accounting prices are discussed, but willingness to pay is never considered an input into accounting prices. The essence of divergences between market and accounting prices is the presence in relevant markets of artificial wedges between demand and supply prices, and the derivation of shadow prices must generally use both.

The first chapter in part 3 is a good survey of formulae for project choice, though Dasgupta and Pearce's denunciation of the benefit-cost ratio seems somewhat harsh, their criticism of the internal rate of return seems too easy, and their example of capital rationing dips by assuming the cost of capital known when its determination is part of the problem. Chapter 8, on risk and uncertainty, is the most successful chapter in the book.

The two case studies are strange and largely inappropriate. The critique of the Roskill Commission study is not related to the rest of the book, and it is in any case a cost-effectiveness study that ignores benefits. The chapter on the Damodar Valley Flood Control Scheme, on the other hand, deals only with benefits, though it fits into the book slightly better.

Most of the studies in the Wolfe volume refer to the point that the role of CBA in decision making will depend on how far the analyst's (implicit) objective function coincides with the decision maker's. Alan Peacock seems to believe that most economists consider CBA a substitute for political decisions and has no difficulty stressing the omnipresence of political judgments, especially in the face of uncertainty. But if Peacock's argument were pushed to the extreme, there would be no role at all for CBA. On the other hand, Peacock seems to confuse political decisions to accept or reject expenditure proposals with the use of CBA as an objective tool for assessing them. Alan Williams, by contrast, views CBA as *assisting*, not *making*, choices. He then presents a persuasive argument for CBA as a scientific procedure which can help to determine what the true political problem is.

Few of the remaining papers represent work at the forefront of current developments. Watson and Mansfield, and Beesley and Dalvi, deal with the evaluation of travel time. Jones arbitrarily prices research and development resources at three to five times their nominal cost. Foster commits the same error as Dasgupta and Pearce commit with respect to the treatment of taxes. The clearest application of CBA is by Walters, on the implications of peak-load pricing for congestion at, and on the advisability of adding runways to, Kennedy International Airport.

On the whole, except for the papers by Williams and Walters, and to a lesser extent that by Heath and Oulton, Wolfe's volume contains little to recommend it. As for Dasgupta and Pearce, only chapters 5–8 can be considered appropriate for an elementary textbook; while chapters 1–3 are written mostly in nontechnical terms, it would require more sophistication to "fill in the gaps" than prospective beginning readers would be likely to command.

DANIEL WISECARVER

Ohio State University

The Crisis in Keynesian Economics. [The Yrjö Jahnsson Lectures.] By JOHN HICKS. Oxford: Blackwell, 1974. Pp. vii+85.

Britain, as the birthplace of John Maynard Keynes, has naturally been the country in which "the Keyesian Revolution" has achieved the most comprehensive dominance of thinking on economic policy. In addition, as James Tobin recently pointed out in a review of a Joan Robinson lecture, British economists tend to believe, more than others, that policy problems are the result of errors of economic theory requiring radical revision of theory. And, finally, "Keynesian economics" in Britain is very much a simplified and bowdlerized *General Theory*, making little if any reference to Keynes's earlier monetary studies and policy writings.

Hence, the current difficulties of British economic policy with the world inflation problem have tended in Britain to be interpreted as a "crisis in Keynesian economics," the subject to which John Hicks devoted his Yrjö Jahnsson Lectures. The lectures themselves, however, are best regarded as the product of Hicks's own thinking over his own lifetime about the successive problems of British monetary theory and policy. As such, however, they suffer on the one hand from a lack of the long historical view Hicks displayed most notably in his *Theory of Economic History*, and on the other hand from undue deference to Keynes and the Cambridge Keynesians.

The three lectures are concerned successively with the three key concepts of Keynesian theory: the multiplier, liquidity preference, and money wage determination (which Keynes himself left an unsolved problem). The multiplier chapter reworks the theory to stress the importance of stocks in permitting output expansion and of what Hicks terms "flexprice" (in contrast to "fixprice") sectors in making quantity expansion have price-inflationary effects. The liquidity chapter is reminiscent of Keynes's "fringe of unsatisfied borrowers" and of the work of Gurley and Shaw on financial intermediation in its main point that monetary policy may be effective in "an overdraft economy" but ineffective in "an auto-economy" (direct self-financing of activity). The chapter ends with the suggestion that inflation is bad because it leads to the reverse of liquidity preference and to hasty and inefficient investment choices. (In fact, recent British inflation has been characterized by the virtual disappearance of shop inventories and stocks, which raises the question of what *does* inflation cause people to invest in. Also, since Hicks produced early in the postwar period an analysis of suppressed inflation under the title "The Empty Economy," it would seem that any kind of inflation, or perhaps full employment by itself, imposes costs on consumers that are simply not measured.)

The third chapter, on wage behavior, is the most controversial and the least satisfying, because in it Hicks attempts to arrive at a balanced view, relying heavily, for the purpose, on the microeconomic, sociology-of-collective-bargaining sections of his first book on the theory of wages. (One of the characteristics of British labor economics, and of that field generally, is its tendency to appeal to microeconomics to explain macroeconomic labor market phenomena.) Hicks distinguishes three actual or potential "Keynesian" positions: (1) Wage behavior is sociological and has nothing to do with full employment, so we ought to have full employment. (2) The Phillips curve exists and offers a trade-off of inflation for employment. Mysteriously, this is identified with demand-pull theory, and Hicks makes the unexpected error of confusing the rate of wage increase with a one-shot increase due to increased aggregate demand. (3) Once wage inflation has started, conventional compromises in achieving "fairness" of wage relativities break down and wage inflation develops its own steam, in the process generating industrial unrest. This is Hicks's own position. But, he claims, this still leaves some choice between less and more inflationary ways of increasing employment.

The book, as will appear from this brief summary, is disappointing. If there is a crisis of Keynesian economic policy, it is scarcely helpful to produce a balanced view of very marginally differentiated Keynesian theories which allows the policymakers some very marginal degree of freedom of choice. And if there is a crisis of Keynesian economic theory, intellectual courage requires much more than tinkering with Keynesian concepts to make them more realistic approximations to perceived reality. If use of Keynes's name is nevertheless essential to academic respectability, it would be better to reread what he actually wrote about inflation in pre-*General Theory* works than to guess what he might reasonably be construed to have said if he had said something in that book.

HARRY G. JOHNSON

University of Chicago

Riches, Class, and Power before the Civil War. By EDWARD PESSEN.
Lexington, Mass.: D. C. Heath Co., 1973. Pp. 378. $7.95.

This book, though the author disavows the purpose of telling a story, is written in clear English, in an engaging style which invests the data offered with almost the spice of a novel or detective story. It is based on "detailed evidence—what nowadays is called 'quantitative data'—gathered in order to subject the thesis of antebellum egalitarianism to the kind of detailed or comprehensive investigation it has hitherto been largely spared" (p. 3). It attacks the thesis of Alexis de Tocqueville, who stated in 1840 that "in a democratic society like that of the United States...fortunes are scanty. ...The equality of conditions [that] gives some resources to all members of the community...also prevents any of them from having resources of any great extent" (de Tocqueville, *Democracy in America*, 2:250, 258; cited by Pessen, p. 9). De Tocqueville's is singled out as the most influential and also the most durable evaluation yet written of Jacksonian America. Not only he but other contemporary extollers of the equalities of the "era of the common man" in the second quarter of the nineteenth century, such as Emerson, James Fenimore Cooper, and European travelers as disparate as Harriet Martineau and Mrs. Frances Trollope, are accused, not of faulty reasoning, but of insights based on slight evidence, and of leaping to the conclusion that the masses dominated society in the New World. The author concludes, after presenting his evidence, that on the contrary the United States of 1825–50 was not an egalitarian society, that there was great concentration of wealth which was becoming greater, that the rich families were rich from generation to generation, that there was not much movement into the class of the wealthy, and that the rich were powerful not only socially but politically. Existence of huge fortunes and concentration of a large portion of wealth in the hands of relatively few are his principal measures of inequality. He compares fortunes and costs of opulent living with those in Europe (p. 30) but concludes: "Until detailed studies are completed on social mobility and distribution of wealth and influence in early nineteenth-century western Europe, it remains impossible to make precise comparisons of the civilizations on the two sides of the Atlantic. ...Certainly the American evidence reduces the disparities so long believed to have prevailed between the two continents" (p. 303). He later states: "The evidence indicates that the egalitarian version of antebellem American society should be discarded" (p. 306).

Author of three previous books and various articles on Jacksonian America, Pessen presents in this book his findings on size of fortunes and distribution of wealth, principally from tax lists. He also offers tabulations of residential location of the wealthy from city directories, showing clearly defined residential patterns of the urban rich and elite, and of occupations of officeholders, coinciding with occupations associated with wealth. He cites some limitations of probate inventories, noted by other scholars and by me, as reasons for their unsuitability for his purpose. I do not concur as to their unsuitability but admit that their use, with appropriate adjustments to compensate for the limitations, would have greatly compounded his task. If someone could undertake a study of a cross section of probate inventories at one date somewhere between 1800 and 1840, even if limited only to some major cities, it might well produce more persuasive evidence for his thesis than Pessen has yet adduced. Acknowledging that there have been other challengers of the egalitarian thesis during the "era of the common man," he says they have "been impressionistic in their methodology; that is their data have been random [*sic*] and inconclusive" (p. 3). He on the other hand aimed to answer questions by "gathering as much data as possible. Evidence was gathered on total populations rather than samples. 'Representative individuals,' even when selected through the most sophisticated statistical techniques, sometimes turn out to be less than representative" (p. 1).

The study includes primarily New York City; Brooklyn, which by 1841 was the seventh largest city in the nation; Boston; and Philadelphia. The "total populations" studied (unless one includes as such occupational counts of members of city councils) are limited to the tax lists for 1828 and 1845 for New York City, 1810 and 1841 for Brooklyn, and local censuses or tax lists for Boston in 1820, 1833, and 1848. The tax lists for Philadelphia proved unusable because property taxes were assessed against occupants rather than owners, but Pessen draws some conclusions regarding the wealthy of Philadelphia from a contemporary list for 1846, which he evaluates as considerably more reliable than the New York City lists of wealthy drawn up in the 1840s by Moses Beach, editor of the *New York Sun*. The latter, previously relied on by several scholars chiefly for their availability, are rather thoroughly discredited in an appendix by comparisons with assessed tax valuations and other checks. Nothing quantitative is offered for other parts of the country or for dates prior to 1810 except for citations of work by other scholars, which are rather completely covered, in footnotes and an extensive bibliography. These include Robert Gallman's estimates of wealth distribution in 1860 for Baltimore, New Orleans, and Saint Louis; K. W. Wheeler on Galveston, Texas; G. Blackburn and S. L. Richards on Michigan; M. B. Katz on Hamilton, Ontario; and R. Doherty on Pittsburgh, in the nineteenth century (p. 41). Comparisons with wealth inequality in the colonial period cite J. T. Main, J. Henretta, A. H. Jones, J. T. Lemon, G. B. Nash, D. W. Kock, and A. Kulikoff as showing that wealth inequality increased greatly during the colonial period, a conclusion which I regard as still open to question. Pessen does not mention B. C. Daniels's article (*Explorations in Economic History* 2 [Winter 1973–74]: 123–35), which shows for some New England towns increasing inequality but for others substantially no change from 1700 to 1776, as revealed in random samples of probate inventories.

Quantitative detail regarding wealth from the tax lists is limited to the first three chapters. Other chapters, in the more conventional vein of literary historical writing, marshal much colorful detail on life styles, social structure, and mobility, with illustrations of balls and drawings rooms, which help cast doubt on the egalitarian thesis and show that at least some Americans lived in as great affluence as the richest of Europe at the time. Material is presented on occupations, marriages, and social life of the wealthy and the passing on of fortunes intact or augmented. Appendix C lists by name New York City's wealthiest 200 in 1828 and 300 in 1845, grouped by size of assessed wealth; Brooklyn's wealthiest 71, based on tax assessments in 1845; Boston's wealthiest 100 in 1833 and 200 in 1848, based on tax assessments; and Philadelphia's reputed wealthiest in 1846, grouped by size of wealth from estimates Pessen considers reasonably reliable, though not verifiable from tax assessments. A few names of women occur on these lists, and sometimes the estate of a person recently deceased is the basis of the tax assessment. The roster includes many names of recognizable influence.

The work on wealth, social structure, and mobility is of interest to a fairly recent breed of social and urban historians as well as to economists, political scientists, and others interested in wealth and its distribution historically. I find the critical appraisal of the underassessments of and lack of disclosure by the rich on the tax lists to be one of the most useful contributions of the work. The inadequacies of the data for these reasons are frankly admitted and are documented as well known at the time, and a rather plausible formula is offered for New York City that the true wealth of the rich, including their real estate outside the city, was at least six times the assessed value. In Boston, assessments were more even-handed, but at half the value, for which Pessen adjusts. He plausibly attributes a large part of corporate taxed wealth to the rich. On page 16 he argues that though wealth shown on tax lists was "too low," there is no reason to believe the relative ranking thereon of rich men to be invalid, or that one would understate appreciably more than another, and that at least the tax assessments form a bedrock minimum appraisal of their wealth. If the understatement was greatest at the upper end of the scale, then the inequality figures Pessen offers are also, if anything, understatements. His only consistent measure of inequality is the proportion of total assessed wealth held by the upper 1 percent of the taxpayers (including their wives and children as part of the upper 1 or indicated percent). He also gives percentages of "population" (?) in various wealth brackets (neither the percentages of "population" nor the wealth brackets remain the same from city to city) holding indicated percentages of total assessed "noncorporate wealth." He scorns use of deciles, Gini coefficients, Schutz coefficients, or Lorenz curves, which measure deviation from perfect equality of wealth distribution. His preference is "to make simple statements of the proportions of wealth owned by different levels of wealthholders. . . . It is more revealing to say that the poorest 90 percent of Philadelphians owned only 10 percent of the wealth in 1860 than to note that the Schutz coefficient of inequality for the same data is .79 (with 0 representing perfect equality and .999 the greatest possible degree of inequality.) Similarly. . . in an instance in which the richest tenth owning better than 40 percent of the wealth and the poorest one half less than 15 percent produces a Gini coefficient of .54" (pp. 5–6). He does not mention possible comparison with a 1962 study by the Federal Reserve Board which gives us a Gini coefficient a little smaller than Robert Gallman's for 1860 (somewhat less inequality in 1962 than in 1860), but for which figures are lacking for percentage of total wealth held by the top 1 or 10 percent. Nor does he mention the work of Lee Soltow on wealth distribution in Wisconsin and in northeast cities, based on the 1860 manuscript census and replete with Gini coefficients, as well as percentages of wealth held by various deciles.

His discussion of price indexes, to compare the value of the dollar in 1840 and 1970 (pp. 17, 29), is naive but reflects a justified intuition that index numbers cannot capture the effect of changing life styles. His training and experience as historian rather than economist in understanding the economic function of wealth is suggested by the following: "Masters of capital, finance, and real estate determined the price and availability of credit, the scale of employment and unemployment, the wages of labor, the price of housing" (p. 249).

Whether Pessen's conclusions, based on four of the older seaboard cities, would hold for newer cities such as Cincinnati, Chicago, Saint Louis, or Milwaukee is not answered by his data. Nor is it clear whether the 1 percent is of the taxpaying or of the total free population, and there is virtually no treatment of the antebellum South. However, the solid and painstaking work on the tax lists for three cities, interpreted against a rich background of collateral information, is a useful contribution to reevaluation of the "era of the common man."

ALICE HANSON JONES

Washington University, Saint Louis